OUTLOOK
Science without borders

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

MIND READING

Visual perception revealed through motor control of cuttlefish skin patterning

PAGES 350 & 361

CONSERVATION BIOLOGY

AN INTERNET OF ANIMALS

One man's dream to track Earth's creatures from space

PAGE 322

AUTUMN READING

BOOK REVIEW SPECIAL

Food fights, designer dogs and bombs in space

PAGE 334

QUANTUM PHYSICS

OUT OF THIS WORLD

Bose–Einstein condensate formed in space

PAGES 351 & 391

NATURE.COM
18 October 2018
Vol. 562, No. 7727

THIS WEEK

EDITORIALS

JOURNALS China takes steps against predatory publications **p.308**

WORLD VIEW Fight the coming pandemic of viral misinformation **p.309**

BALLISTIC Russia investigates cause of rocket failure **p.312**

Progress on antibiotic resistance

Clinicians, companies and researchers have come together to suggest ways to break the deadlock on finding better ways to prescribe antibiotics.

Accounts differ on exactly how long Alexander Fleming was away from his London laboratory on holiday before he discovered the impact of penicillin. It might have been two weeks; it might have been four. But we do know it was long enough for the famous stray *Penicillium* mould to develop and wipe out colonies of bacteria on his discarded petri plates. Some things — the growth of microbial life among them — simply can't be rushed.

Today, that creates a problem. Getting an infectious agent identified and the best antibiotic prescribed within a typical eight-hour working day is close to impossible. It generally takes several days, and sometimes more. And the longer it takes to start treatment, the more time the infection has to take hold.

It is understandable, then, that physicians would rather not wait. One way to speed things up is to take a best guess at the diagnosis and throw a broad-spectrum antibiotic at it (one that works against a number of bacteria). That approach can save lives, but it brings its own problems. The World Health Organization predicts that, without urgent action, the spread of antibiotic-resistant bacteria will lead to a resurgence in deaths from minor injuries and previously benign infections.

A good place to focus urgent action is the delay between a person becoming ill and receiving effective treatment. Shortening that time would reduce unnecessary prescribing, minimize the spread of resistance and, most importantly, give people the best chance of recovering.

Speeding up that process will require major advances in what microbiologists call antimicrobial susceptibility testing. Conventionally, this testing can be broken down into two steps. First, laboratories culture and identify the infectious agent. And second, they show which antibiotic is most likely to be effective.

In theory, the technology exists to hasten both of those stages. Advances in genomics mean that rapid DNA sequencing can identify bacteria within hours. It can also quickly and accurately detect antibiotic resistance and susceptibility for tuberculosis (The CRyPTIC Consortium and the 100,000 Genomes Project *N. Engl. J. Med.* **379**, 1403–1415; 2018).

Developed further, this and other technologies could deliver results within an hour of a sample being taken from a patient. That would be a game-changer — but it has not yet happened. Why?

Talk to the people involved — physicians, researchers, testing labs, regulators and commercial firms among them — and all will offer their own reasons. One result of such discussion is published this week— a consensus statement that seeks to find common ground on defining the obstacles and recommending ways to overcome them (A. van Belkum *et al. Nature Rev. Microbiol.* https://doi.org/10.1038/s41579-018-0098-9; 2018).

The statement is signed by specialists who represent organizations from the French diagnostics company bioMériux to the European Commission's Joint Programming Initiative on Antimicrobial Resistance, which coordinates national research programmes. It's a landmark effort and a triumph of cooperation and communication for the community. Now the hard work really begins: addressing the issues in the roadmap.

One challenge is regulation. Regions and countries tend to have their own requirements and processes to approve the marketing of new diagnostics and to validate them after market release, so developers have a gargantuan task to meet all the different demands. Here, communication, harmonization and standardization are needed: policymakers need to sit together and agree on a common set of rules.

> ***"Communication, harmonization and standardization are needed."***

Another issue is how institutions collect and compile information about resistant strains and the usefulness of antibiotics. If the information were made available in real time and more samples analysed, the statement says, then a "smart antibiogram" could be developed to guide treatment. That could bring down the time to treatment — and time saved is lives saved.

Perhaps the biggest obstacle is cost. Current diagnostic tests might be slow, but they are cheap. Modern diagnostics tend to be more expensive to develop and use. That could change: rising antibiotic resistance could shake up clinical practice and the health-care market so fundamentally that most of today's treatments and diagnostics become obsolete. In that case, what we now regard as too much of an investment will seem comparatively cheap. The world must not wait for such dire circumstances. Policymakers repeatedly say that action is needed on antibiotic resistance. The community has responded with a way forward. ■

Dandelion clocked

The flight of a dandelion seed shows how wonderful discoveries can lie right in front of us.

The English poet and artist William Blake was no fan of the reductionism of Isaac Newton. True discovery, and therefore knowledge, Blake insisted in his poem 'Auguries of Innocence', was to be found in the everyday, where a world could be seen in a grain of sand and "heaven in a wild flower".

Today, we know of exotic states of matter that can slow the vast speed of light to a mere sprint. And astronomers have spotted more than 3,800 planets in more than 2,800 distant stellar systems: a staggering rate of discovery, given that the first confirmed detection of a planet orbiting another star similar to the Sun was as recent as 1995.

None of this should blind us to the fact that — as Blake suggested — some of the most surprising discoveries come from the world

of the familiar. No one has visited an exoplanet, but most people know what a dandelion looks like. This flower (*Taraxacum officinale*) is found worldwide. And, as many a child discovers to their delight, when a dandelion sets seed, the flower (actually, hundreds of tiny florets) turns into a mass of seeds known as a dandelion clock. Each seed is suspended from a parachute-like stalk — easily released by a puff of breath.

The parachute is a bunch of bristles called a pappus. Each pappus carries around 100 filaments, each attached to a central point, rather like the head of a chimney sweep's brush. Just like a parachute, it increases aerodynamic drag, slowing the descent of each seed and allowing it, once aloft, to be wafted kilometres from the parent plant. So much we know.

Here's the surprising part — the mechanism of this dispersal was unknown until now. As researchers write in *Nature* this week (C. Cummins *et al. Nature* https://doi.org/10.1038/s41586-018-0604-2; 2018), the bristles are arranged so that when the pappus falls, air flows between them and creates a low-pressure vortex, like a smoke ring. This vortex travels above the pappus and yet is not attached to it, an invisible yet faithful familiar that generates lift and prolongs the seed's descent.

The key lies not in the bristles of the pappus, but in the spaces between them. If projected on to a disc, the bristles together occupy just under 10% of the pappus's area, and yet create four times the drag that would be generated by a solid disc of the same radius. The study shows that air currents entrained by each bristle interact with pockets of air held by its neighbours, creating maximum drag for minimum expenditure of mass. The pappus's porosity — a measure of the proportion of air that it lets pass — determines the shape and nature of the low-pressure vortex.

All falling objects, from feathers to cannon balls, create turbulence in their wake. But it takes a rare combination of size, mass, shape and, crucially, porosity for the pappus to generate this vortex ring. Size is also particularly important, because from the point of view of something as small as a pappus, the air is appreciably viscous. At such a scale, a parachute consisting of a bunch of bristles is as effective as the aerofoil found in larger seeds that disperse from taller plants — such as the winged seeds of the maple. In the same way, the tiniest insects do not fly with solid wings, but swim through the air using 'paddles' made of bristles.

"Some of the most surprising discoveries come from the world of the familiar."

It's an example of how evolution can produce ingenious solutions to the most finicky problems, such as seed dispersal. There are many things unknown that are smaller than atoms, or larger than galaxies, or billions of years away in time. But there are secrets held by things that we take for granted — things on a human or near-human scale — that seem all the more precious for it. Heaven in a wild flower, even. ■

On the list

Compilations of academic journals to use or avoid need transparent criteria.

This January, China was reported as overtaking the United States to become the largest producer of scientific papers. There is one major caveat, however, which consoles those who worry about China's rise and worries those who cheer for it: a lot of those Chinese publications are of poor quality.

Over the past few years, China has taken steps to show that it is serious about fixing this problem. Officials are censuring individual scientists to deter them from fraudulent activity and are upping the pressure on the universities that might try to protect them. In May, China set its sights on a more ambitious target — predatory journals, those that put no effort into vetting papers and exist only to collect money that scientists pay to get their research published. Officials announced punishments for scientists who publish work in journals that the government feels are not good enough. The Chinese government has not yet announced which journals it intends to blacklist. But institutions such as universities and hospitals are already establishing their own lists of journals to avoid — to the vexation of some researchers.

China is not the first to make such an effort. Hunting down poor-quality and shady journals has become a mission for some librarians and governments. Most famously, Jeffrey Beall, a librarian at the University of Colorado Denver, started a list in 2008 of journals he said were dubious, which grew to more than 1,000 titles.

But creating such lists is not easy. Most scientists and scientific policy-makers would agree that it's good to condemn predatory journals. But it can be difficult to distinguish them from ones that operate in good faith, but which might have published some poor-quality or fraudulent research because of short cuts in editorial decision-making due to lack of resources, because scientists deceived them, because of lapses in judgement — or because people just make mistakes.

Listing such journals would risk denigrating some good research. That's why, although many researchers supported Beall, others criticized his list for a lack of clear standards. The list was taken down in January 2017, but there have been new incarnations.

Some say a better approach is to produce lists of approved journals. That does solve some problems. For example, in logistical terms, it is easier to maintain. Instead of trying to track down every newly emerging predatory journal, the burden is on the journals to prove themselves. It does not generate the same stigma as bans, and thus allows reputations to be redeemed. The journal *Tumor Biology* was behind one of the most egregious research scandals to hit China — the retraction of 107 papers by Chinese authors in 2016. But it now has a new publisher and, since January 2018, a new editor-in-chief who is hoping for "a new chapter" for the journal. In August 2018, *Tumor Biology* was listed in the Directory of Open Access Journals, a vote of confidence from a website that catalogues high-quality publications. But *Tumor Biology* still appears on the emerging Chinese lists.

Whether it is fair to continue snubbing such a journal brings up a central question about the grading process: are the criteria for listing clear, transparent and consistently applied? This is the only way that the system can be fair to all parties — scientists who want to publish good papers, journals that want to communicate solid science and governments that want to ensure their funding is being spent wisely.

At present, the criteria for a journal to appear on a Chinese blacklist are not clear. This understandably leads some researchers to wonder why their research should be devalued just because it was published in the same journal as some poor-quality research. Word of mouth is even creating informal bans that could overturn the genuine achievements of a researcher.

Establishing and agreeing on such criteria is not a simple task. The analytics company Cabells in Beaumont, Texas, maintains a blacklist and lists suspicious signs that mark a suspect publication, such as the inclusion of fictional or dead editors, and poor spelling. Criteria to appear on an approved list might be more practical: as a minimum, journals should list their profit or non-profit status clearly, list editors who are aware they are editors, use basic technology to detect plagiarism, and carry out due diligence to ensure that, if reviewers suggested by the author are used, they exist, are competent in the field, and are the ones being contacted.

Publishers have an obligation to maintain standards so that scientists and governments can rely on them in evaluating research and achievements. But to do so, they need feedback when those who depend on them believe they are falling short. ■

CLARISSA SIMAS

The biggest pandemic risk? Viral misinformation

A century after the world's worst flu epidemic, rapid spread of misinformation is undermining trust in vaccines crucial to public health, warns **Heidi Larson**.

A hundred years ago this month, the death rate from the 1918 influenza was at its peak. An estimated 500 million people were infected over the course of the pandemic; between 50 million and 100 million died, around 3% of the global population at the time.

A century on, advances in vaccines have made massive outbreaks of flu — and measles, rubella, diphtheria and polio — rare. But people still discount their risks of disease. Few realize that flu and its complications caused an estimated 80,000 deaths in the United States alone this past winter, mainly in the elderly and infirm. Of the 183 children whose deaths were confirmed as flu-related, 80% had not been vaccinated that season, according to the US Centers for Disease Control and Prevention.

I predict that the next major outbreak — whether of a highly fatal strain of influenza or something else — will not be due to a lack of preventive technologies. Instead, emotional contagion, digitally enabled, could erode trust in vaccines so much as to render them moot. The deluge of conflicting information, misinformation and manipulated information on social media should be recognized as a global public-health threat.

So, what is to be done? The Vaccine Confidence Project, which I direct, works to detect early signals of rumours and scares about vaccines, and so to address them before they snowball. The international team comprises experts in anthropology, epidemiology, statistics, political science and more. We monitor news and social media, and we survey attitudes. We have also developed a Vaccine Confidence Index, similar to a consumer-confidence index, to track attitudes.

Emotions around vaccines are volatile, making vigilance and monitoring crucial for effective public outreach. In 2016, our project identified Europe as the region with the highest scepticism around vaccine safety (H. J. Larson *et al. EBioMedicine* **12**, 295–301; 2016). The European Union commissioned us to re-run the survey this summer; results will be released this month. In the Philippines, confidence in vaccine safety dropped from 82% in 2015 to 21% in 2018 (H. J. Larson *et al. Hum. Vaccines Immunother.* https://doi.org/10.1080/21645515.2018.1522468; 2018), after legitimate concerns arose about new dengue vaccines. Immunization rates for established vaccines for tetanus, polio, tetanus and more also plummeted.

We have found that it is useful to categorize misinformation into several levels. Among the most damaging is bad science: people with medical credentials stoking overblown or unfounded fears. The canonical example is the 1998 publication by infamous former physician Andrew Wakefield purporting to show a link between autism and the measles, mumps and rubella (MMR) vaccine. Despite having his licence revoked and his work retracted, Wakefield persists in campaigning against the vaccine. Expert consensus alleges that his efforts have contributed to persistent vaccine anxieties and refusals, including a 2017 measles outbreak in Minnesota. Had Wakefield been disciplined and his article retracted 12 months after publication rather than 12 years, we might not be remarking that this year marks the twentieth anniversary of its publication.

The second-most-dangerous category includes those who see anti-vaccine debates as a financial opportunity for selling books, services, or other products. (Wakefield, who maintains that financial concerns have not affected his research and that he has been unfairly vilified, gave paid testimony against the vaccine and filed a patent that allegedly stood to become more valuable were the vaccine to be discredited.)

The next tier of damaging misinformation comes from those who see anti-vaccine debates as a political opportunity, a wedge with which to polarize society. Multiple reports this year found that Russian trolls and bots used emotional, angry language to spread misinformation and exacerbate the divisions between those for and against vaccines (see D. A. Broniatowski *et al. Am. J. Pub. Health* **108**, 1378–1384; 2018).

EMOTIONS AROUND VACCINES ARE VOLATILE, MAKING VIGILANCE CRUCIAL FOR PUBLIC OUTREACH.

Next are 'super-spreaders', who propagate misinformation through social media to like-minded vaccine-questioners. A common claim is that suspected adverse reactions to vaccines (typically coincidences) are confirmed reactions. Finally, there is misunderstood or inadequate information that might be circulating generally.

Targeted social media can combat misinformation. Both Denmark and Ireland faced groups broadcasting testimonies on social media and television news of young girls alleged to have been harmed by human papillomavirus (HPV) vaccination. In Denmark, national immunization rates fell from over 90% in 2000 to under 20% in 2005.

In response, Danish public-health officials emphasized the risk of disease, and promoted stories of people who had lost wives and mothers to cervical cancer. They also created a Facebook page for answering parents' questions. Ireland's social-media efforts used similar tactics to rebuild HPV-vaccine confidence; numbers for 2018 show an increase of 6% for vaccine uptake from 2017.

No single strategy works for all types of misinformation, particularly among those who are already sceptical. Educational materials and resources are important, but limited; health officials and educational campaigns often fall short because they craft messages based on what they want to promote, without addressing existing perceptions. Dialogue matters. Strategies must include listening and engagement.

We have to get better at this: if a strain as deadly as the 1918 influenza emerges and people's hesitancy to get vaccinated remains at the level it is today, a debilitating and fatal disease will spread. ■

Heidi J. Larson *is professor of anthropology, risk and decision science at the London School of Hygiene & Tropical Medicine.*
e-mail: heidi.larson@lshtm.ac.uk

SEVEN DAYS *The news in brief*

EVENTS

Climate win

A court of appeal in The Hague has upheld a precedent-setting judgment that forces the Dutch government to step up its efforts to curb greenhouse-gas emissions in the Netherlands. In 2015, a district court in The Hague had ruled in favour of the Urgenda Foundation, a climate-change group that filed the lawsuit on behalf of 886 Dutch citizens. The foundation asked for more-stringent government action to protect the low-lying country from the harmful effects of climate change. The government appealed against the verdict, arguing that courts have no right to take decisions on this matter. The appeal judges disagreed. On 8 October, the court of appeal confirmed that the government must take measures to cut domestic greenhouse-gas emissions to at least 25% below 1990 levels by 2020. The court cites the state's legal duty of care for its citizens, which is enshrined in the European Convention on Human Rights. Similar court cases are ongoing in several countries, including the United States, Belgium, Norway and Ireland.

TASS VIA GETTY

Astronauts escape Soyuz rocket crash

The Russian space agency, Roscosmos, is investigating why a Soyuz MS-10 malfunctioned on 11 October just after takeoff from the Baikonur Cosmodrome in Kazakhstan. The rocket was carrying Russian cosmonaut Alexey Ovchinin and US astronaut Nick Hague to the International Space Station. An alarm notified the crew of a problem with the rocket's booster roughly 90 seconds into the flight. An automated system immediately detached the crew capsule from the rocket, and the astronauts began a ballistic descent — a steep, rapid dive to Earth. The capsule landed about 500 kilometres northeast of the launch site in Dzhezhazgan, Kazakhstan, and search-and-rescue teams took the crew to the Gagarin Cosmonaut Training Center near Moscow for medical attention. Roscosmos and NASA officials say that both crew members are doing well; Russia's deputy prime minister, Yury Borisov, tweeted that Roscosmos will suspend crewed missions until it can guarantee the safety of launches.

POLICY

Looking down

Australia should invest in a "downward-looking telescope" that could study mineral resources as far down as 300 kilometres beneath Earth's surface. The proposal is one of several in an Australian Academy of Science report on the future of geosciences in the country, released on 15 October. A geological 'telescope' would consist of a network of geophysical remote sensors and geochemical sampling programmes. Geoscientist Sue O'Reilly at Macquarie University, who chaired the committee that wrote the report, says this approach is unique. Australia introduced the idea of digitally integrating information across geophysics, geochemistry, geology and tectonics, she says. The report highlights the need for Australia to find new sources of minerals that will be needed in vast amounts for a future based on low-carbon renewable energy. Copper and cobalt are essential components of electric cars, and rare earths are used in solar cells.

Alcohol warnings

Public-health officials in Australia and New Zealand have welcomed the two governments' 11 October decision that alcohol products must carry standardized labels warning about the risks of drinking during pregnancy. Such labelling is currently voluntary. Research suggests that labels can increase awareness about the effects of drinking during pregnancy, such as the risk of fetal alcohol spectrum disorder, which is the most common preventable cause of non-genetic intellectual disability in Australia. But warnings are unlikely to change behaviour by themselves.

PEOPLE

Ribosome pioneer

Thomas Steitz, a Nobel-prizewinning biochemist, died of pancreatic cancer on 9 October. He was 78 years

ERIK DE CASTRO/REUTERS

old. Steitz shared the 2009 Nobel Prize in Chemistry for determining part of the complex molecular structure and function of cellular machines known as ribosomes, which read RNA messages encoded by DNA, translate them into amino-acid sequences and then assemble those amino acids into proteins. The discovery revealed that many widely used antibiotics kill bacteria by binding to the cells' ribosomes and disrupting their function. Steitz worked at Yale University in New Haven, Connecticut, for 48 years. His wife, Joan Steitz, is also a biochemist at Yale and won the 2018 Lasker-Koshland Special Achievement Award in Medical Science for her own work on RNA biology.

FACILITIES

Rice-diversity boon

A major global repository of rice biodiversity has secured permanent funding. The International Rice Research Institute gene bank, based in Los Baños, the Philippines, harbours some 136,000 varieties of rice and its wild relatives — resources that scientists can study and that breeders can use to develop new kinds of rice. On 12 October, the Crop Trust, a non-governmental

organization in Bonn, Germany, that supports food security and crop diversity, announced funding of US$1.4 million a year "in perpetuity" to support the gene bank. Rice is a staple food for almost half the world's population (**pictured**, golden rice), and new varieties could help to maintain crop production in the face of increased drought, flooding and other manifestations of global climate change.

AI investment

The Massachusetts Institute of Technology (MIT) in Cambridge has committed up to US$1 billion to set up a new college for artificial intelligence and computer science. An initial $350-million donation by Stephen Schwarzman, chief executive of investment firm Blackstone in New York City, will help to create 50 new faculty positions and construct a building for the college, according to a 15 October announcement. Half of the posts will be joint appointments with other campus departments to enable research and education across disciplines, including the ethics of ground-breaking technologies. The goal is to educate students so that they are proficient in computing as well as their main field of study. MIT has already raised a further $300 million and is working to raise the remainder. The college is scheduled to open in September 2019.

SPACE

Hayabusa2 delay

The mothership of the Japanese probe Hayabusa2 will make its first touchdown on the Ryugu asteroid in January, instead of this month as originally planned. On 14 October, the Japan Aerospace Exploration Agency said it had determined that the asteroid's surface is rougher than expected, and so has decided to take more time to plan the landing. This part of the mission — which will collect a sample of the asteroid to be brought back to Earth — is the most important, but also the riskiest. Hayabusa2, which carries several smaller probes, launched in late 2014 and reached Ryugu in June this year, aiming to return samples to Earth by 2020. Hayabusa2 is hovering over the space rock at varying altitudes and has already deployed three small landers, which have sent back images and data from the surface.

RESEARCH

Stem-cell studies

Harvard Medical School and Brigham and Women's Hospital (BWH) in Boston, Massachusetts, have called on journal editors to retract 31 papers co-authored by Piero Anversa, a former BWH lab director, after their investigations concluded that the studies contained falsified or fabricated data. The papers focused on cardiac stem cells, which Anversa claimed could be activated to allow damaged heart tissue to regenerate itself. In 2017, Partners HealthCare System, which runs BWH, agreed to repay US$10 million to the US government to resolve allegations that Anversa's lab had fraudulently obtained federal grant funding. The lab closed in 2015. Earlier this year, Anversa was selected to be an expert adviser at Italy's National Institute of Health in Rome. He withdrew his candidature after objections by Italian scientists.

SOURCE: ARXIV/KEVIN FLAHERTY

TREND WATCH

Women with astronomy PhDs are leaving the field before landing a faculty job at a rate three to four times higher than that for men, a study of crowdsourced hiring data in the United States has revealed.

Kevin Flaherty, an astronomer at Williams College in Williamstown, Massachusetts, collected hiring data for 2010 to 2017 from the Astrophysics Jobs Rumor Mill, a website where astronomers can anonymously aggregate information about the status of open fellowships and faculty positions. He found 245 reports of tenure-track faculty positions in US universities that went to 157 men and 88 women. By Google-searching the year each astronomer received their PhD, he found that women landed their faculty jobs, on average, a year sooner than did men (K. Flaherty https://arxiv.org/abs/1810.01511; 2018).

To try to explain this, Flaherty created a model of the labour pool with hires and departures, and ran three scenarios: one, that more women were receiving astronomy PhDs over time; two, that female astronomers were more likely than men to be hired; and three, that women were leaving the field at a higher rate than men. The third model fit the data best.

ASTRONOMY HIRING GAP

Female astronomers in the United States are hired in academia, on average, a year earlier than men. This difference appears to stem from women leaving the field three to four times faster than men. Because more women leave, those hired quickly are overrepresented in the data, and so the average time for those who do get hired is shorter.

RICK BOWMER/AP/SHUTTERSTOCK

The long reach of law enforcement could now extend into our DNA.

CONSUMER GENETICS

Privacy concerns over DNA used for crime investigation

Almost all Americans of European descent could soon be traced using relatives' DNA.

BY EWEN CALLAWAY

Genetic sleuthing techniques that led to the arrest of a suspect in the infamous Golden State Killer case this year are set to become vastly more powerful, suggest two papers published last week[1,2].

They conclude that it could soon be possible to search crime-scene DNA for links to nearly all Americans of European descent, while vastly expanding the reach of an existing forensic database. The results raise urgent privacy issues. "It's important to have this discussion early on," says Yaniv Erlich, chief scientific officer at consumer genetics firm MyHeritage in Yehuda, Israel, who led one of the studies[1].

From the mid 1970s to the late 1980s, a string of burglaries, sexual assaults and murders in California were attributed to an unknown person dubbed the Golden State Killer or the East Area Rapist. The case went cold, but in April 2018, police arrested Joseph James DeAngelo. He was identified as a suspect, in part, by matching crime-scene DNA to genetic profiles posted by his distant relatives on the genetic-genealogy website GEDmatch, which allows people to upload data from consumer genetic companies and to search for relatives. Between April and August 2018, more than a dozen cases have been solved using this technique.

Erlich and his team set out to measure the reach of the method, known as long-range familial search. They analysed anonymized DNA profiles from 1.28 million MyHeritage customers. Like similar firms, the company allows customers to search for relatives who share DNA inherited from a common ancestor.

The researchers found that 60% of MyHeritage customers had a third cousin ▶

or closer relative in its database. Searches of 30 randomly selected GEDmatch profiles found a similar rate of relative matching.

But such databases could identify many more people who aren't in them. DeAngelo was not on GEDmatch; detectives found him using profiles of his third cousins. Erlich's team estimates that a database containing genetic profiles of 3 million Americans of European descent could enable the identification of 90% of this demographic using public genealogy records. GEDmatch is growing by 1,000–2,000 profiles per day, says co-administrator Curtis Rogers, and should hit 3 million in the next few years.

To see whether they could track down people not in the database, the researchers attempted to identify an anonymous woman from Utah who had made her DNA public as part of the 1000 Genomes project. They uploaded her profile to GEDmatch and searched for distant cousins. Of the people who had enough DNA in common with her to suggest that they shared an ancestor in the past few generations, two also had enough public genealogical information to narrow the search. After a day spent ruling out hundreds of descendants, the team identified the Utah woman. (She is not named in the paper and the researchers made no attempt to contact her.)

SPOTTING INFORMATION

DeAngelo was identified only because crime-scene DNA had been preserved. This allowed forensic scientists to apply the approach now used in consumer genetics testing and many biomedical studies: sequencing hundreds of thousands of DNA variants, or single-nucleotide polymorphisms (SNPs), across the genome.

For the past few decades, by contrast, most crime-scene DNA has been analysed using the sequences of more than a dozen 'short tandem repeats'. The FBI's Combined DNA Index System (CODIS) holds more than 13 million such profiles. These allow forensic scientists to determine an individual's genetic signature, but are poorly suited to matching relatives, says Noah Rosenberg, a population geneticist at Stanford University in California. To circumvent this, Rosenberg's team developed a computational method to cross-match CODIS profiles with a close relative's SNP profile.

The team identified an anonymous woman who had made her DNA public.

Simulations suggested that about one-third of people genotyped using short tandem repeats could be correctly matched to a first-degree relative genotyped with SNPs[2]. This could allow investigators who are unable to generate SNP profiles from crime-scene material to look for matches to CODIS profiles in databases such as GEDmatch, and vice versa, Rosenberg says.

Forensic genealogical investigations similar to the Golden State Killer case are set to grow. The lack of regulation for such searches is striking, says Rori Rohlfs, a statistical geneticist at San Francisco State University in California. However, some rules do exist: in California, for example, law-enforcement forensic databases can be used to find relatives only in cases of serious crimes where there is a risk to public safety, and the genealogical investigative team must be distinct from local detectives on a case.

Erlich says that consumer genetics companies could include digital signatures with the data files people can download, allowing GEDmatch to differentiate them from crime-scene profiles uploaded by investigators, and shield consumers from searches. Rogers says that GEDmatch has no plans to limit law-enforcement access — after the Golden State Killer case emerged, the site updated its terms of service to explicitly warn users that investigators could use it — and he worries that regulating use will interfere with the site's purpose: helping people find relatives. "I don't think anyone's privacy is being violated," he says. "People should be able to control their own DNA and not the government."

Colleen Fitzpatrick, co-executive director of the DNA Doe Project in Sebastopol, California, which has used familial searching to help solve a number of missing-person cases, says the information gleaned from these searches isn't so different from other leads — and therefore shouldn't be treated differently. "Just about anything we do in life reveals information about others," she says. ■

1. Erlich, Y., Shor, T., Pe'er, I. & Carmi, S. *Science* https://doi.org/10.1126/science.aau4832 (2018).
2. Kim, J., Edge, M. D., Algee-Hewitt, B. F. B., Li, J. Z. & Rosenberg, N. A. *Cell* https://doi.org/10.1016/j.cell.2018.09.008 (2018).

FUNDING

Argentina's scientists struggle as peso slips

Inflation and currency devaluation have hobbled research.

BY MICHELE CATANZARO

Juan Pablo Paz's plans for a new cold-atom laboratory have slowly eroded over the past two years. Paz, a physicist at the University of Buenos Aires, won a US$1.1-million grant in February 2017 to set up the facility. But the money, awarded by the Inter-American Development Bank, was transferred to Paz through an Argentinian government agency that paid him in pesos.

As Argentina's currency weakened, so did Paz's buying power. When the physicist won his grant, $1 cost 16 pesos. "Now it costs 38," he says. "By the time I got the money, I was able to buy just a part of the equipment."

Paz, who is looking for money to cover the last 40% of his lab's start-up costs, is one of many researchers who say that Argentina's worsening financial woes are hurting their research. The slipping peso makes it harder to purchase equipment from abroad, while rapidly increasing inflation has crushed scientists' budgets and salaries at home.

Researchers have also struggled under austerity measures adopted by the government in 2014 and intensified in June by a financing agreement that Argentina signed with the International Monetary Fund.

"The science and technology system of Argentina is collapsing," a group of high-profile scientists, including Paz, wrote in an open letter published late last month. More than 1,000 foreign scholars or Argentinian scientists working abroad — including several Nobel laureates — have endorsed the message.

Argentina's total science spending increased tenfold between 2003 and 2015, reaching the equivalent of $3.96 billion. Along the way, in 2007, the country established a dedicated science ministry. Still, Argentina spends much less of its gross domestic product (GDP) on research than does South America's leader, Brazil. The slice of GDP that Argentina devotes to science peaked at 0.63% in 2012, when Brazil spent 1.13%.

And recent years have seen a reversal in fortune for research overall. The government's science outlay fell by almost 40% between 2015 and 2018 when measured in US dollars, and the share of Argentina's budget devoted to research has fallen from 1.69% in 2008 to 1.23% in 2017. The picture is set to grow grimmer next year: the budget proposed by Argentina's president, Mauricio Macri, includes further cuts to science.

PINCHED PURSE

The government's belt-tightening has drastically reduced the average value of awards made by ANPCyT, Argentina's main granting agency for science and technology. The country has also suspended its contributions to several international research projects.

The situation is also dire at CONICET, Argentina's national research council. The council, whose budget stood at $681 million

A professor teaches students in Buenos Aires during a protest for better wages and funding.

EITAN ABRAMOVICH/AFP/GETTY

last year, is now spending 90% of its money on salaries and scholarships, leaving little for research, says Fernando Stefani, a physicist at the University of Buenos Aires. "There are research centres that cannot pay for illumination or gas. Their lab rats and cell lines are dying," he says. "It's a dramatic situation."

Scientists also complain of delays in payments from CONICET and ANPCyT. "We are in October, and we have been transferred less than 40% of our annual budget," says Andrea Gamarnik, a virologist at the Leloir Institute, a life-sciences research centre in Buenos Aires.

Officials with the government say they hope to ease the pain, but keeping up with the peso's slide has been difficult. "We were not able to change the budget immediately," says Jorge Aguado, secretary of planning and policy at Argentina's science secretariat. "We understand the concerns, but we are committed to extend the budget, in order to maintain projects and purchases."

Mario Albornoz, coordinator of the Ibero-american Network of Science and Technology Indicators (RICYT), says the situation in Argentina is part of a broader trend. "Almost all Latin American countries, including Brazil and Mexico, are cutting their science budgets, for macroeconomic reasons," says Albornoz, whose group tracks statistics related to research in the Americas. "This government has made many mistakes, but it's not true that it wants to destroy science."

But that does not satisfy many researchers, who see science as vital to Argentina's future. "What will we live off in 30 years?" says Stefani. "Past-century technology and agricultural activities will not be enough." ■

INDONESIA

Clash over tsunami access

Scientists say red tape imposed by Indonesian government is delaying research.

BY QUIRIN SCHIERMEIER

Two weeks after an earthquake and subsequent tsunami killed more than 2,000 people on the Indonesian island of Sulawesi, some foreign researchers say that red tape is slowing down or preventing investigative work of the devastated coastlines. But the Indonesian government says that it has sped up the time it takes to process permits for researchers in the wake of the tsunami, and that the requirements it imposes on international researchers have been in place for years.

"It is absolutely important for us to go to the field to survey the correct locations," says tsunami researcher Philip Liu, vice-president for research and technology at the National University of Singapore. "But when I asked for a permit, I understood that it might take months." As a result, Liu decided not to research the area after all.

Meanwhile, an international reconnaissance team led by Costas Synolakis, a tsunami researcher at the University of Southern California in Los Angeles, had rushed to Singapore a week after the tsunami hit, hoping to get to Indonesia. But the researchers learnt that they must submit detailed survey plans and research proposals that include local collaborators. They say this rule was not enforced before, and fear that it might delay the planned survey by several weeks, time they can ill afford. "Disaster surveys need to mobilize in the first few days after the disaster, before the data needed to better understand the event is permanently eradicated," says Synolakis. The Sulawesi events are of particular interest to scientists in Southern California and the Mediterranean, where active tectonic faults close to the coast could likewise trigger unpredictably large tsunamis, says Synolakis, who is still in Singapore awaiting his research permit, although some of his team have returned home.

But Sadjuga, the head of the team at the ▶

HARIANDI HAFID/SOPA IMAGES/LIGHTROCKET/GETTY

International researchers are waiting to be granted access to study the aftermath of the tsunami.

research ministry that grants research permits, says that international researchers have been required to apply for permits and report findings to their Indonesian partners for a "long time". "It has been the normal procedure in Indonesia," he says.

Sadjuga says the government understands the importance of timely data collection at the site. "That is why we are currently speeding up the research permit process." He says that it normally takes researchers 14–28 days to gain a research permit, but for teams wanting to visit the city of Palu, where the tusnami hit, the ministry is trying to give them out within 7 days.

WORK IN PROGRESS

Two international teams, one from South Korea and another from the United States, have applied for permits, Sadjuga says. "We gave a research permit to a Korean team on October 10. The US proposal has not been granted a permit as the applicants have not completed all the requirements," he says.

Synolakis, who is behind the US proposal, says it will take at least another week to meet all the requirements.

A few Japanese researchers have collected data in the disaster area along with the local survey team. Taro Arikawa of Chuo University in Tokyo presented the preliminary results of their survey at a 10–11 October tsunami workshop in Singapore.

It is still unclear exactly what kind of underwater disturbance triggered the tsunami. Tide-gauge data, and reported tsunami height and arrival time suggest a source near the entrance to the Bay of Palu, says Liu, who convened the meeting. "It could be a submarine landslide triggered by the earthquake, or it could have been generated by sudden subsidence of the sea floor," he says. Arikawa plans to return to the disaster region this week to collect more data. He promised the meeting he would report back to colleagues who are unable to do field work in the area. "As long as the tsunami community exchanges ideas and information openly it does not matter so much whether I can get in," says Liu. "But there are so many different ideas and so much to do. Allowing only a few people to go in might mean that a lot of fresh evidence and information will be lost."

J. C. Gaillard, a geographer at the University of Auckland in New Zealand, says Indonesia is right to take control of post-disaster research. "No one knows and understands the context and local concerns, including research needed to enhance disaster risk-reduction policy and practice, better than the Indonesians," he says. "This does not mean that foreign researchers should be excluded."

The Indonesian government has submitted a draft law to parliament that proposes tougher penalties for foreign researchers who break existing regulations. ■

MICHAEL GOTTSCHALK/PHOTOTHEK/GETTY

Plant biology is at the centre of a long-running saga over scientific misconduct.

MISCONDUCT

Biologist cleared

French national research council absolves one lab leader of misconduct, and holds another researcher responsible.

BY DECLAN BUTLER

France's national research council has ruled that one of its plant biologists committed misconduct through manipulation of published figures, including data fabrication, but it cleared another researcher whom it had heavily sanctioned in 2015.

The ruling adds some clarity and closure to the long-running saga — although the cleared researcher, Olivier Voinnet, is now raising questions over how the French research agency, CNRS, handled its initial investigation.

The CNRS announced its conclusions on 3 October, following a fresh inquiry that it led — with the participation of the Swiss Federal Institute of Technology Zurich (ETH Zurich) — into five articles published by researchers at a now-defunct lab at the CNRS Institute of Plant Molecular Biology in Strasbourg, France. The lab was renowned for its work on a gene-silencing technique called RNA interference.

The CNRS and ETH Zurich each drew their own conclusions about their respective staff members, on the basis of the inquiry's report.

ETH Zurich said last month that the inquiry found "severe" and "intentional" manipulation of research figures. However, it said that Voinnet, a former leader of the Strasbourg lab and a prominent CNRS scientist who has been on secondment to ETH Zurich since 2010, "did not perform, order or scientifically endorse such manipulation". But ETH Zurich concluded that, as former group leader and a co-author of four of the papers, Voinnet bore overall management responsibility. The institution therefore extended until 2023 a probation it had implemented after its 2015 investigation, including monitoring his publication activity and assigning him a mentor.

The CNRS has now reached a similar conclusion with respect to Voinnet. *Nature* has obtained a copy of the conclusions of a 10 July meeting of the CNRS disciplinary committee, which advises CNRS management on appropriate sanctions. The document states that after studying the CNRS–ETH Zurich report, and after interviewing the inquiry committee's president and Voinnet, the committee found no evidence of serious wrongdoing by Voinnet — and voted 7 to 0 (with one abstention) in favour of no sanctions against him.

In a statement released on 3 October, the CNRS reiterated that its disciplinary committee had found no evidence that Voinnet was responsible for unethical manipulations of figures or data in the investigated papers. But, like ETH Zurich, it said that as a former head of the group, Voinnet bore some management responsibility, and so gave him a reprimand

that will stay on his record for three years, and which is 'category 1', the least serious sanction on the four-tier scale used in the civil service.

The finding contrasts with the 2015 CNRS investigation, which found Voinnet guilty of research misconduct and suspended him from the agency for 2 years — a category 3 sanction. At the time, the agency found no evidence of data fabrication, but said the intentional manipulation of figures breached ethical standards.

The CNRS has also now said that, in its joint report with ETH Zurich, the institutes concluded that another former researcher at the laboratory, Patrice Dunoyer, committed misconduct in the form of figure manipulations — and in corrections to the manipulated papers — including data fabrication. The CNRS said that Dunoyer would receive the category 2 sanction of a demotion, a more severe punishment than Voinnet's but still relatively low.

In 2015, Dunoyer had received a 12-month exclusion from the CNRS, without pay, for scientific misconduct, with 11 of those months served as a suspended sentence. Alain Schuhl, deputy director-general in charge of scientific affairs at the CNRS, told *Nature* that this suspended sentence will now kick into effect.

The latest version of the CNRS's official bulletin, published on 10 October, confirms the charges and sanctions against Dunoyer, and the reprimand on Voinnet. Yet the first version of the bulletin, published on 9 October, made no mention of the minor sanction the CNRS gave to Voinnet, which a CNRS spokesperson attributes to a "computing bug".

Dunoyer, who the CNRS statement says is on temporary assignment at the secretariat general of South Province of New Caledonia, his place of birth, has not replied to *Nature*'s requests for comment. Loic Dusseau, Dunoyer's lawyer, told *Nature* that Dunoyer asked him to consider whether to appeal the CNRS ruling, and says that Dunoyer feels the ruling is unfair and questionable.

> ***"A reprimand is what I only should have got in 2015 — and not a two-year suspension."***

DRIVING FORCE

The latest probe was instigated at the initiative of Voinnet, according to ETH Zurich, after he raised the possibility of more-serious misconduct than had been found in 2015.

Voinnet says that the "CNRS's reprimand is perfectly in line with the conclusions of ETH Zurich last month exonerating me". But he takes issue with the agency's 2015 ruling: "A reprimand is what I only should have got in 2015 — and not a two-year suspension."

Schuhl declined to comment on the seeming reversal of responsibilities in the CNRS's latest conclusions compared with 2015. He said that the matter of its 2015 investigation is closed.

Questions about papers co-authored by Voinnet were first raised in January 2015 on the PubPeer website, which allows anonymous commenting about research articles. The CNRS announced in April 2015 that it had set up a commission to investigate the affair — and in July that year, it announced the original sanctions against Voinnet. At the time, its official bulletin referred only to an inquiry commissioned in early 2015, and to an 8 June 2015 meeting of CNRS's disciplinary committee. *Nature* has obtained a copy of the confidential report of the inquiry, which comprised three CNRS scientists and two from other French research organizations who, between 29 January and 2 February 2015, inspected the relevant articles and interviewed several people, including Voinnet and Dunoyer. The report makes no mention of lab notebooks or raw data, unlike the latest investigation, and runs to four pages.

Voinnet hopes that the recent investigation will lift "the cloud of suspicion that has hung over many other members of the lab". But he told *Nature* that he now intends to take administrative legal action against the CNRS to challenge the grounds for his 2015 sanction. In response, the CNRS spokesperson said that the latest investigation has no bearing on the sanctions pronounced in 2015. ■

CLIMATE IMPACTS

Trouble brewing for beer prices

Extreme weather will cut barley yields and drive up drink costs, say researchers — but the increase could encourage more people to pay attention to climate change.

BY MATTHEW WARREN

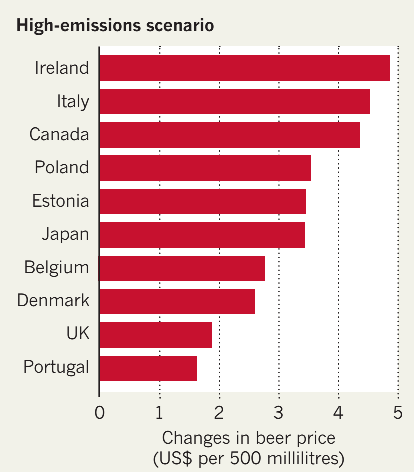
Extreme weather caused by climate change can have devastating effects — and it turns out that not even beer is safe.

More-frequent droughts and heat waves in the twenty-first century will reduce global production of barley, finds a study published on 15 October (W. Xie *et al. Nature Plants* http://doi.org/cvtm; 2018). In turn, it shows, this will decrease the supply of beer, drive up prices and cut consumption, even under best-case climate-change scenarios.

Studies have previously explored how climate change will affect staple foods and luxury goods. But nobody has considered how beer will fare, says Dabo Guan, a climate-change economist at the University of East Anglia in Norwich, UK. It might seem trivial to consider beer production, but Guan hopes that helping people to understand how climate change could affect their daily lives will motivate them to take action. "What I'm trying to emphasize here is that climate change will impact people's lifestyle," he says. If people "want to drink beer when we watch football, then we have to do something", he says.

CLIMATE'S TOLL ON BEER

Models show that during years of drought and heat waves driven by climate change, the global supply of barley — and therefore beer — will decrease and prices will rise.



SOURCE: W. XIE *ET AL. NATURE PLANTS* HTTP://DOI.ORG/CVTM (2018).

The team began by examining the chances of major droughts and heat waves occurring in barley-growing regions on all six inhabited continents between 2010 and 2099. They considered four futures, based on different emissions scenarios, from low to high emissions throughout the century.

In each case, extreme weather was likely to become more frequent in barley-growing regions compared with the number of similar events recorded in the late twentieth and early twenty-first centuries. In the best-case scenario, the chance of extreme weather increased by a modest 4%, but the worst case saw a rise of 31%. The researchers then simulated the effect of these droughts and heat waves on barley production by using software to model crop growth and yield on the basis of weather and other variables. They found that, globally, ▶

extreme weather would reduce barley yield by between 3% and 17%. Some countries fared better than others: tropical areas such as Central and South America were hit badly, but crop yields increased in some temperate areas, including northern China and the United States. But this was not enough to offset the global decrease.

Finally, Guan and his colleagues fed these changes in barley yield into an existing economic model to look at how reduced barley production would affect pricing and consumption of beer. In the worst-case scenario, the reduced barley supply would result in doubling of prices and a 16% decrease in beer consumption in the years of extreme-weather events (see 'Climate's toll on beer').

Klaus Hubacek, an ecological economist at the University of Maryland in College Park, says that the study does a good job of combining climate, agriculture and economics models. He wonders how other alcohol crops might be affected, and whether beer drinkers might switch to cider or other alcoholic drinks.

But worries about beer pale in comparison to projections of how climate change could harm food security generally, says David Reay, a climate-change scientist at the University of Edinburgh, UK. "The effect on beer is going to be the least of our worries," he says, especially in the worst-case climate scenarios. Reay worries this message could be diluted in studies such as Guan's, which concentrate on luxury items.

"I think in that kind of future, I probably will need a beer, because it will be pretty bad," Reay says. ■

SPACE

Mercury probes ready to begin seven-year journey

BepiColombo, a joint Europe–Japan mission, is only second ever mission to the planet.

BY DAVIDE CASTELVECCHI

A European rocket is ready to launch the most ambitious mission ever to Mercury, Earth's once-neglected sibling in the Solar System. The €1.6-billion (US$1.85-billion) expedition, carrying 2 robotic orbiters, ranks among the most expensive missions undertaken by the European Space Agency (ESA), and includes Japan's largest contribution yet to an international collaboration in space.

If all goes according to schedule, BepiColombo will lift off in the late hours of 19 October from the Kourou spaceport in French Guiana, atop an Ariane 5 heavy-launch vehicle, to embark on a seven-year journey to Mercury. When it gets there, it will release two probes into the planet's orbit: the Mercury Planetary Orbiter (MPO), built by the European Space Agency (ESA), and the Mercury Magnetospheric Orbiter, nicknamed MIO and built by the Japan Aerospace Exploration Agency (JAXA).

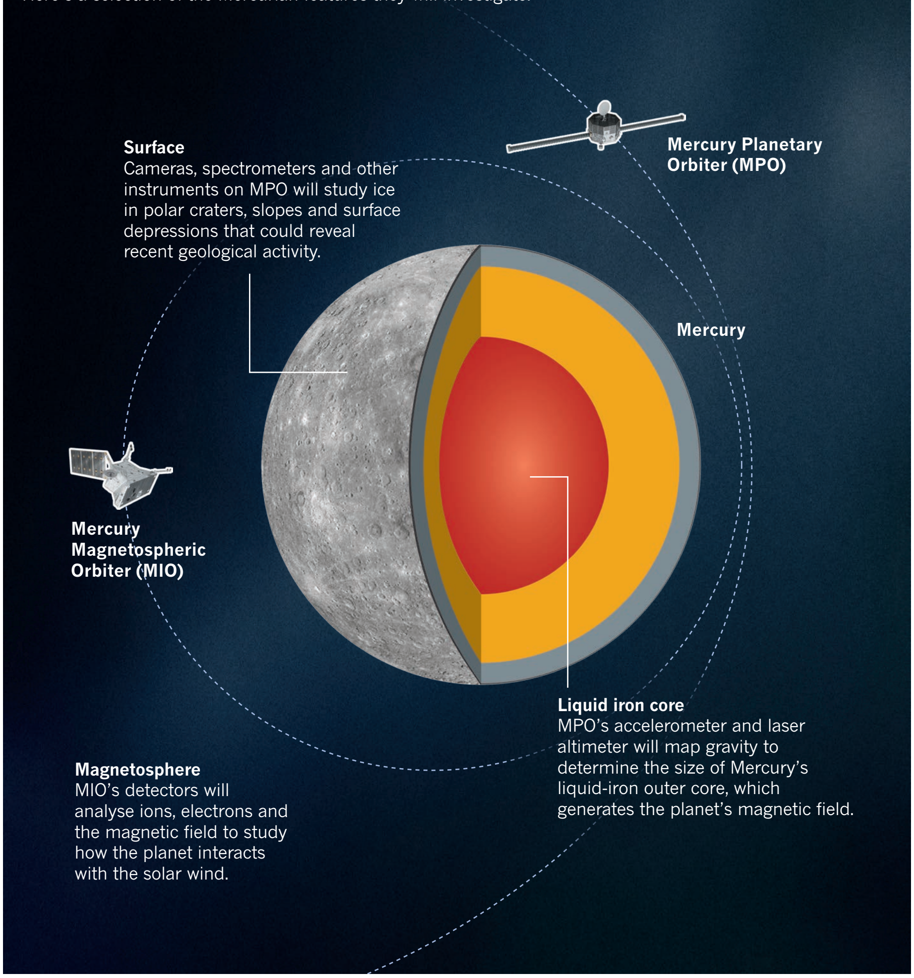
The orbiters will investigate the mysteries of the innermost, smallest planet of the Solar System (see 'Journey to Mercury'). Mercury was once thought to be a static, boring place. But in recent years, it has revealed many surprises, from its unusual magnetic field to water-ice deposits in some of its craters.

BepiColombo was first conceived in the 1990s and has had a long, complicated gestation, says Johannes Benkhoff, overall project scientist and a planetary physicist at ESA in Noordwijk, the Netherlands. "It's a great moment," says Benkhoff, who has worked on BepiColombo for nearly 15 years. "Now it's becoming real."

Mercury is deep in the Sun's gravitational well, which makes reaching it a challenge. To get there, a spacecraft has to lose much of

JOURNEY TO MERCURY

The BepiColombo mission, which should reach Mercury in 2025, carries two orbiters armed with a host of instruments that will probe the mysterious planet's chemistry, geology and magnetosphere. Here's a selection of the Mercurian features they will investigate.



SOURCE: ESA

the momentum it gains from Earth's orbital motion, so that it can fall towards the Sun. But it must also avoid overshooting. Because of this, it takes eight times as much energy — and a lot more time — to travel to Mercury than to Mars. BepiColombo will use advanced, solar-powered ionic thrusters combined with gravitational assists from a total of nine fly-bys of Earth, Venus and Mercury itself.

Moreover, sunlight is ten times more intense at Mercury than it is in outer space near Earth, and the planet's nearly atmosphere-free surface reaches temperatures of 400 °C. All these factors have made Mercury the least explored of the four planets of the inner Solar System: the only other probe to have entered its orbit was NASA's MESSENGER, which studied the planet from 2011 to 2015. An older NASA Probe, Mariner 10, made several fly-bys of Mercury in 1974 without entering its orbit.

The new mission is named after Giuseppe 'Bepi' Colombo, the late Italian scientist who studied Mercury and conceived of Mariner 10's gravitational-assist trajectory. MPO and MIO will each have their own science priorities. MIO will focus on the environment around Mercury — especially the magnetic field and its interaction with the solar wind. MPO will mainly scan and map the planet's surface, using instruments that can analyse most of the electromagnetic spectrum as well as neutrons that can reveal the chemical composition of the planet's crust. It will also study the gravitational field — and through that, the planet's unusually large iron core — and test some subtle predictions of Albert Einstein's general theory of relativity.

BIG OPPORTUNITIES

"The thing that's really exciting about MPO is [its] low-orbit altitude," says Nancy Chabot, a planetary scientist at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland, who was a leading scientist for MESSENGER. This will enable MPO to map the entire surface at high resolution. It might even spot the 16-metre-wide crater that MESSENGER created when it dropped onto the surface at the end of its mission — potentially leaving interesting layers of rock exposed, Chabot says.

Chabot and her collaborators have found compelling evidence for the presence of ice deposits in the permanently shaded areas of some craters near the north pole. Further studies of those craters — including some that might exist at the south pole — could motivate a future mission, possibly with a lander. "Getting down to the surface is the next step," says Chabot, who is part of a working group that will try to make the scientific case for such a mission.

Meanwhile, MIO will spin continuously to get a full-sky view of Mercury's magnetosphere and the particles that wind around it, says MIO project scientist Go Murakami. "Our particle sensors can cover almost [the entire] field of view," says Murakami, a planetary scientist at JAXA's Institute of Space and Astronautical Science in Sagamihara. The intense solar wind around Mercury might be comparable to the stellar wind around planets that closely orbit relatively cool red dwarfs — the most common stars in the Milky Way. So, studying Mercury could help scientists to understand what conditions might be conducive to life on extrasolar planets, Murakami says. The BepiColombo probes are designed to last at least two years in orbit, although their mission could stretch a bit longer. But sooner or later, the heat will catch up with them, Benkhoff says. "Our time is limited. Mercury is a harsh environment." ■

CORRECTION

The News story 'Brazil's presidential election could savage its science' (*Nature* **562**, 171–172; 2018) incorrectly stated that Jair Bolsonaro had proposed eliminating the science ministry and reorganizing it under the agriculture ministry. The candidate has discussed efforts to decentralize federal science programmes in Brazil, but it's unclear how he would do so.

The internet of animals

Martin Wikelski has spent 17 years getting an antenna into space to track animals around the world. That's the first step in his plan to revolutionize biology.

BY ANDREW CURRY

On a Wednesday afternoon in August, biologist Martin Wikelski watched helplessly as 17 years of his professional life was about to go to waste — because of a cable mix-up.

Some 400 kilometres above Earth, cosmonaut Oleg Artemyev was fumbling with an electrical connection while Wikelski monitored the operation from a command centre in Moscow. The cosmonaut was floating outside the International Space Station (ISS) and trying to manoeuvre the thick, stiff gloves of his spacesuit to join two unmatched cables — something that was never going to succeed. As the ISS hurtled through space at more than 27,000 kilometres per hour, the cosmonaut made no headway with his task.

Wikelski, the director of the Max Planck Institute for Ornithology in Radolfzell, Germany, was in Moscow to witness the culmination of a quest that had absorbed much of his career — and had threatened to derail it at times. Artemyev was installing a three-metre-long antenna on the outside of the space station as part of a system designed to track wildlife on Earth. This project would enable scientists to spy on animals from space for the first time, including ones so tiny that they can't carry current satellite-tracking devices.

CHRISTIAN ZIEGLER

In the long run, Wikelski hopes the system will connect so many individuals — from elephants and warblers to baby sea turtles — that it could create an 'internet of animals'. It could use the movements and habits of wild creatures to reveal patterns in much the same way that mobile-phone apps pinpoint traffic and illuminate people's social networks. But first, the cosmonaut had to plug in the antenna on the ISS.

The 15 August spacewalk had been agonizing to watch from the

Martin Wikelski studies a bat in Zambia as part of a tagging expedition.

start. In an auxiliary room at the Moscow command centre, Wikelski winced as he saw Artemyev grapple with the 120-kilogram antenna and manhandle it out of the station's airlock, knocking the fragile receiver into the exterior of the spacecraft. Hours went by as Artemyev and his fellow cosmonaut Sergei Prokopyev crawled across the ISS, painstakingly stringing cables around its cluttered exterior.

Finally, as the Russians struggled to connect the antenna to a power source, Wikelski was plagued with doubts. Had his team somehow made a horrible mistake with the wiring that was causing problems for the cosmonauts? Was all this work destined to fail?

He huddled with a small team of engineers from the German Space Agency and his own institute, watching Artemyev attempt again and again to make the connection, while checking the blueprints to make sure he hadn't let the wrong cable somehow slip through. Wikelski knew the long-delayed project wouldn't get another chance.

Finally, after minutes that seemed like hours, Artemyev located the right cable end and plugged the antenna into the station's power supply. Engineers in Moscow flicked on the antenna's computers. One by one, its systems came online. As the last line on the screen moved from red to green, Wikelski could finally relax.

Days later, back at the Max Planck institute he's led for the past decade, the 53-year-old researcher stood up in front of his gathered staff. "Nominal operation should start in early November," Wikelski announced exuberantly, raising a glass in a toast. "Now we are spacemen!"

FLYING HIGH

Over the past few decades, tracking wildlife using radio collars and GPS (Global Positioning System) transmitters has changed the way that researchers understand the behaviour of the animal kingdom. Using tags that communicate through satellite, mobile-phone and radio technology, scientists can follow everything from whales in the open ocean to jaguars beneath deep jungle cover.

But the long-range movements of most of the world's species remain invisible to researchers. Animals that weigh less than 100 grams can't safely carry the smallest available satellite tags. That puts 75% of all bird and mammal species — and all insects — off limits to this kind of tracking. And the tags themselves cost thousands of dollars apiece, making wide-scale deployment a pricey proposition.

Wikelski hopes to change all that with his project, called ICARUS (International Cooperation for Animal Research Using Space), which goes well beyond the single ISS antenna. Within ten years, he foresees a network of satellites devoted to following hundreds of thousands of animals in real time.

"If you don't understand what they're doing in the wild, you don't understand biology."

The internet of animals envisioned by Wikelski would be able to answer questions that researchers didn't even know they had. ICARUS could, he says, illuminate why migratory bird and bat species are disappearing, and map the spread of pathogens such as bird flu and Ebola. It could even provide early warnings of pest outbreaks and, possibly, earthquakes. "By elevating our viewpoint into space and looking down on the globe, it changes the approach to ecology," he says.

It has taken Wikelski a large part of his career to get ICARUS aloft. He came up with the idea of a radio receiver mounted on the ISS back in 2001. Ever since, he has been pitching it to funders and other biologists, while waiting for the technology — and everyone else — to catch up with his vision.

It hasn't been easy. NASA officials at the Johnson Space Center, where ISS mission operations are based, laughed him out of Houston. He eventually won a spot for ICARUS on the Russian ISS module, but spent years worrying that international political strife would sink that chance. He focused so much of his effort on ICARUS that, at one point, his position at the Max Planck Institute for Ornithology was at risk. Even after he'd grown ICARUS into a project involving dozens of people, partners in five countries and eight major funding institutions, Wikelski told colleagues in January that he was on the verge of giving up, unwilling to lose more years of his research career to a quixotic dream.

But he didn't, or couldn't, let it go. "Mere mortals would not have achieved this," says ornithologist David Winkler at Cornell University in Ithaca, New York. "He deserves a ton of credit for putting so much of his life into this. It's been a tremendous investment."

And the real work, it turns out, could just be getting started.

UP IN THE AIR

As a boy, Wikelski spent hours staring up at swallows and house martins sheltering under the eaves of his family's barn in the Bavarian countryside. When a teacher told him that the tiny birds roamed as far as South Africa each year, it sparked his imagination. In his teens, he photographed birds and trained as a bird bander, fixing tiny strips of metal to fledgling swallows and marvelling when they returned years later — sometimes to the same nests. "When you band an animal and it comes back after a global journey, it's really incredible," he says.

The discovery began a lifelong quest to get as close as possible to migrating animals. During his mandatory military service in the early 1980s, Wikelski volunteered for early shifts as a transport driver, waking at 5 a.m. to finish his working day by early afternoon. As soon as he was off-duty, Wikelski headed for the heights of the Bavarian Alps, hang-glider in tow. "I had the chance to hang-glide every day for a year," he says.

Suspended high above the ground, he was able to feel the air currents that carried birds and bats aloft. "It was transformative," Wikelski says. "I wanted to understand what birds were doing, and you can't understand if you don't do it yourself."

Wikelski earned a behavioural ecology PhD in Germany, then headed for a postdoc in the United States and quickly on to the Smithsonian Tropical Research Institute in Panama, before taking a position at the University of Illinois in Urbana-Champaign. On the flat plains of the US Midwest, he traded his hang-glider for a beaten-up Oldsmobile with purple velour seats and an antenna protruding from the roof. His graduate students referred to the contraption as the Batmobile.

It might not have looked like much, but when it came to understanding migrating birds, the Batmobile was state-of-the-art. Researchers had used similar set-ups since the 1960s, when pioneering US biologist William Cochran used tiny radio tags to track migrating songbirds such as the Swainson's thrush (*Catharus ustulatus*). The transmitters were light enough for the small songbirds to carry, but the trade-off was a short range: Cochran, and later Wikelski, had to follow within a few kilometres of the birds to pick up the radio signals.

Because Swainson's thrushes fly at night and can move at up to 112 kilometres per hour when the winds are right, tracking the birds takes the skills of a rally-car driver and the endurance of a marathon runner. "The thing takes off anywhere between [dusk] and 2 a.m., and as soon as that beep changes you start driving like crazy because you don't want to lose that bird," Wikelski says. The Batmobile offered the necessary acceleration. Speeding along at 3 a.m., the unusual car with its peculiar antenna would get stopped by local police two or three times each night.

In 2004, Wikelski and Cochran teamed up with biologist Henrik Mouritsen to work out how the thrushes navigate after dark. They placed captive birds in magnetized cages to artificially reorient them, then released them. Racing behind them night after night — including a 1,100-kilometre odyssey across the US Great Plains to chase a bird smaller than a clenched fist — they were able to show that the birds

A. GERST, NASA/ESA

In a marathon spacewalk in August 2018, cosmonauts attached the ICARUS antenna to the International Space Station.

used a combination of magnetic sensing and light cues to calibrate their flight path[1].

Wikelski has since adapted the technique to track ever-smaller creatures. He has successfully mounted radio tags to cicadas, dragonflies and even bumblebees, and continues to follow radio-tagged birds and bats across Europe. He's learnt that long-distance migration is much more common than was thought, and that some insects fly for kilometres to find food. The work shows that migration is much cheaper for animals in terms of energy output than researchers ever imagined: bats and birds float on updraughts, butterflies 'swim' in the airstream, and some birds have the same heart rates during flight as they do while sitting, says Wikelski.

The work cemented his belief that tracking the natural movements of animals is key to unlocking their behaviour. "If you don't understand what they're doing in the wild, you don't understand biology," Wikelski says.

But it was apparent that chasing after animals on the ground one at a time was always going to yield limited results. "If you really want to understand the world, you have to do it from above," Wikelski says.

TRACKING IN THE JUNGLE

Wikelski first tried tracking animals from above in the late 1990s, on a 16-square-kilometre island off the coast of Panama called Barro Colorado. He and biologist Roland Kays, now at North Carolina State University in Raleigh, wanted to follow jungle creatures such as jaguars, agoutis and sloths as they moved through the thick forest. But GPS was in its early days, and the forest's thick canopy thwarted tags equipped with the technology.

Instead, Wikelski and Kays adapted radio tags and built a network of 7 radio towers, each more than 40 metres tall, to triangulate signals from animals on the move. The software they devised to process and store their data became the basis for a system called Movebank, which lets biologists around the world analyse and share movement-tracking data.

An internet of animals could answer questions that researchers didn't even know they had.

The system was launched in 2007 and collected its billionth data point in September, providing the basis for hundreds of scientific publications. Some of its information on animal movements is also accessible to the public through a mobile-phone app called Animal Tracker.

By linking ICARUS and Movebank, Wikelski hopes to create a powerful tool that both researchers and the public can use. Standing in the shadow of the Radolfzell castle near his office, Wikelski pulls out his smartphone and taps on the Animal Tracker app. He calls up a type of duck called a Eurasian wigeon (*Anas penelope*) that researchers had nicknamed Guillaume. The duck wears a tracking tag that connects with mobile-phone networks, which shows he's been bobbing on a pond in Kazakhstan for the past two weeks.

A button on the app lets users easily scroll back in time. Wikelski traces Guillaume's zigzagging trail back across Europe to the outskirts of Amsterdam, where the duck was captured and tagged six months earlier. In an age in which people watch live streams of eagle's nests and obsess over individual animals while thousands more disappear unnoticed, Wikelski thinks such tracking data are a way to personalize conservation. "Finally we have a way to live with a wild pet. We can finally understand how difficult and dangerous it is. You can see that duck on your local pond just got back from Russia," he says. For a moment, the boy who watched swallows set off for South Africa shines through.

Tracking animals, Wikelski argues, is a way to "Cecilize" conservation. Cecil was a charismatic male African lion (*Panthera leo*) and was one of the most photographed — and beloved — animals in Hwange National Park in Zimbabwe. In 2015, Cecil was hunted and killed by a US dentist outside the park, but data from the lion's tracking collar revealed he was lured outside the protected area. Cecil's death sparked an international uproar and triggered calls for a ban on trophy hunting.

Wikelski sees opportunities to capture the same kind of interest about other wildlife problems, such as the rapid decline in European songbird

populations. He would like to raise awareness by tracking what happens to them. "We're missing 420 million songbirds, and no one cares," he says. "One Cecil the shrike changes everything."

Wikelski's experience tracking animals in Panama and the United States prompted him to wonder about a better method. Why bother with radio towers or chasing birds in a car or aeroplane when you could put your receiver in space, where it would be able to pick up signals from around the world, regardless of geography?

The idea seemed so powerful and obvious that, when he first proposed it in 2001, he assumed it would be an easy sell. "I thought, 'In three years we'll have it on the space station,'" he says. Instead, it took years for Wikelski, by then an assistant professor at Princeton University in New Jersey, to even get an appointment at NASA. When he went to the agency's Houston space centre in 2004, the earnest young German biologist was shuffled from office to office with no success. Birds and bats, he quickly learnt, weren't on NASA's radar.

Wikelski refused to let the idea go. "After NASA said, 'It's never going to fly,' I called it ICARUS," after the doomed character from Greek mythology who plummeted into the sea after flying too close to the Sun. He began reaching out to colleagues around the world, gathering interest and examples of how lighter, cheaper satellite tags could make a huge difference in everything from large-mammal conservation to sea-turtle research.

The result was a 2008 white paper that listed 32 possible applications of the technology and carried signatures from dozens of prominent biologists worldwide[2]. Within the field, the idea of lighter, cheaper satellite tags was a hit. "The challenge was convincing the space people it was worth it," says Kays, one of the project's founding partners. "What he's been doing for the past ten years is talking to the space people."

Around the time he released the white paper, Wikelski was offered one of the most coveted positions in science: a Max Planck directorship. In 2008, he took over the Max Planck Institute for Ornithology, moving from Princeton to the institute, which was housed in a countryside castle near Konstanz.

The transition, he says, was rocky. Wikelski shut down the institute's bird-banding operation — the same one through which he had first learnt to band birds as a teenager — earning him the enmity of many traditional bird-watchers in Germany. And even Wikelski's renowned endurance was tested as he struggled to run an institute, teach at the nearby University of Konstanz and continue pressing forward with ICARUS.

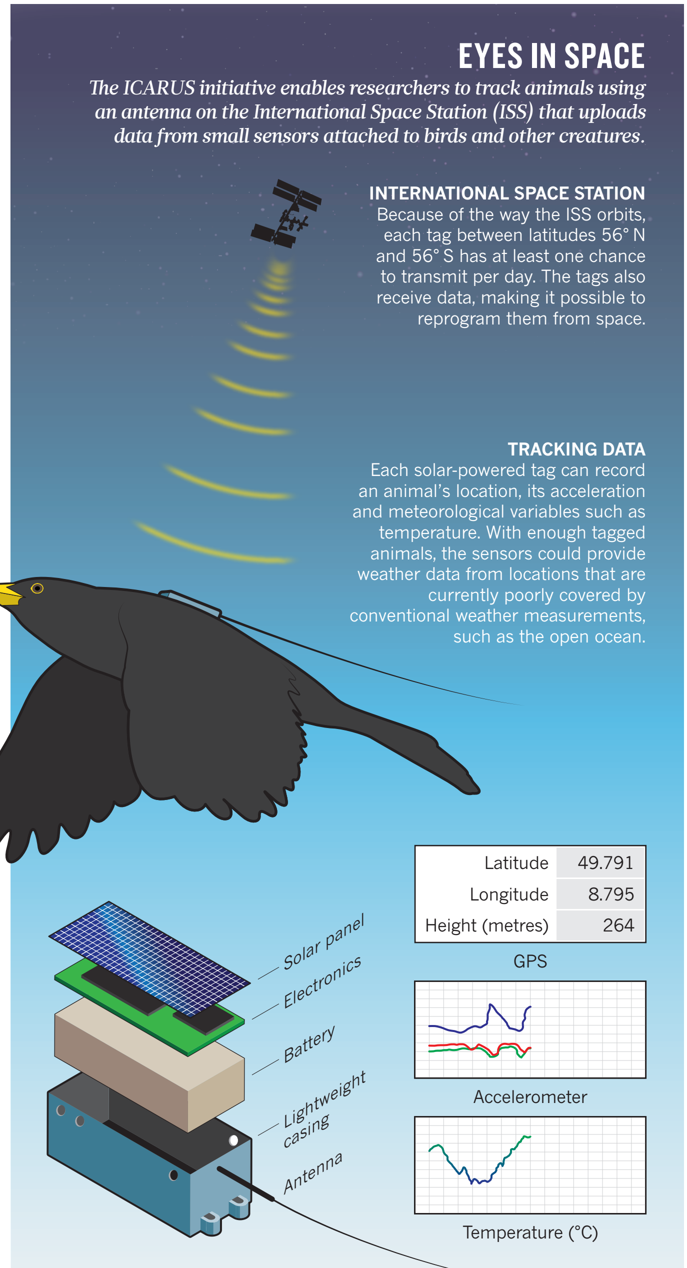
A few years after he took over, Wikelski says, external reviewers gave him a failing grade as a director, citing ICARUS and Movebank as distractions. He was at risk of funding cuts or of losing his directorship. Herbert Jaeckle, the vice-president of the Max Planck Society at the time, says Wikelski persuaded the society to trust him and his ICARUS plans. "He was almost fanatical with respect to this idea," Jaeckle says. "We were convinced he was going to do it, and we were right."

ICARUS started to gain momentum after the head of the German Space Agency (DLR) heard a pitch from Wikelski and told him to apply for funding. The DLR was more enthusiastic than NASA had been, but still struggled to work out where an animal-tracking project might fit into a space agency's priorities, says Johannes Weppler, the project manager at the DLR who is now in charge of the ICARUS programme.

Eventually, in 2012, the DLR agreed to fund ICARUS as a technical experiment, and spent more than €27 million (then US$35 million) to develop, test and build the ICARUS antenna now on the ISS. Russia, the project's other national partner, provided the room on the station, the crew to install it and the rocket to carry it up to space. The launch was tentatively scheduled for 2015.

As relations between Russia and the West grew rocky, the launch was delayed again and again. "At a certain point, I wondered if it would be cheaper to just give up," Wikelski says. By the beginning of 2018, he vowed: "If the antenna's not going up in February, another few months and we quit."

Suddenly, things began to move. In February, Wikelski was at the Russian space centre in Baikonur, Kazakhstan, to watch ICARUS's antenna array leave Earth. And in mid-August, he watched from Moscow as the system sent its first signals back to Earth.



IN FLIGHT

Just a few days later, Wikelski drives his convertible Fiat 500 onto a dirt airstrip on the edge of Konstanz. Bounding up the steps of the two-storey control tower, he greets the lone air-traffic controller on duty with a hug. Wikelski spends a lot of time here — he's at the controls of the institute's Cessna aeroplane for at least 200 hours each year.

Pulling open a hangar door, Wikelski pushes the small red-and-white prop plane outside. After a pre-flight check and a bumpy take-off over a rutted grass field, he banks north above the blue waters of Lake Constance. As Wikelski heads for the forested hills to the north, he turns on an antenna mounted on one of the wing struts and balances a tablet computer on his lap. "We're simulating the ISS, basically," he says.

Somewhere down below, he explains over the roar of the engine, are five blackbirds wearing some of the first ICARUS tags deployed in the wild. The 5-gram tags each contain a thermometer, accelerometer and GPS receiver, plus a transmitter that can send a signal into space and a solar-charged battery to power it all (see 'Eyes in space').

CHRISTIAN ZIEGLER

Biologist Martin Wikelski frequently pilots his institute's Cessna aeroplane to track tagged animals.

The device's small size, Wikelski says, makes a huge difference. Tags that weigh more than 3% of an animal's body weight have the potential to alter its behaviour and threaten its survival. That's why the vast majority of animals are off limits to standard tags that use mobile-phone or satellite technology. Once the ICARUS system is fully operational later this autumn, the tags will transmit 220-byte strings of data at a time up to the ISS. That's the equivalent of 20 GPS positions, enough to provide a sketch of an animal's movements on any given day.

Thanks to a solar-powered battery, ICARUS tags can theoretically last as long as the animal that's carrying them — and can be retrieved and reused. The tags also include a memory chip that can store up to 500 megabytes of data — enough to record an animal's travels, movement and energy expenditure over a lifetime.

For faster data transmission, researchers can download information from the tag using a handheld radio device if they can get close enough — anywhere from a few hundred metres to 15 kilometres, depending on geography and vegetation. "That's very exciting," says Emily Shepard, a specialist in bird energetics at Swansea University, UK.

As the plane flies 1,000 metres above the wooded hills around the Radolfzell institute, information begins to appear on Wikelski's tablet: the location of the blackbird tags, how much battery power is left and when they last communicated with the receiver. Each byte sent skyward provides details about the birds' habits.

Colleagues say that the array of sensors on the tags offers researchers the opportunity to answer thorny but crucial questions about animal behaviour. They could, for example, explore why birds choose certain flyways, by combining accelerometer data on the number of times they flap their wings, and their GPS positions, with wind speed and precipitation records. Scientists could use all of that to compare how much energy it would cost for a bird to take one route instead of another[3].

Monitoring bird migrations is only the beginning. To sell the idea of a €27-million animal-tracking antenna to policymakers, Wikelski leant heavily on its potential benefits for humankind. Tracking the airspeed and temperature of thousands of birds, he argues, amounts to creating a low-cost, distributed weather-monitoring system across the globe[4].

"In the future, we'll use every animal that flies as a meteorological drone," he says. "To measure the temperature in the middle of the Pacific at 20-metres altitude is impossible, but birds do it all the time."

He doesn't plan to stop with the weather. One of Wikelski's most daring ideas depends on deploying ICARUS tags in areas that are prone to seismic activity. Folktales are full of animals that can predict seismic events. So Wikelski thought it might be possible to create an earthquake early-warning system by putting tags with accelerometers on animals in seismically active areas.

To test the idea, in 2012 Wikelski tagged semi-feral goats that roam the slopes of Mount Etna, an active volcano in Sicily, Italy, with data-logger tags that let him analyse their movements after he recovered the tags. Over the course of several years, he observed the goats moving around much more during the 4 to 6 hours before major eruptions than after the events. "If you have a distributed network of goats on the mountain and they all go crazy on some nights, it's pretty simple," Wikelski says.

For all the potential of an internet of animals, Wikelski recognizes that colleagues still have reservations about creating such a system. If it is deployed on the scale Wikelski imagines, unretrieved tags might amount to high-tech litter in some of the world's least-accessible places. And no matter how light the tags might be, catching creatures and placing trackers on them subjects the animals to increased risk. "We need to be asking, just because we can tag something, should we?" says Shepard. "As the cost decreases and the access increases, it's going to be something important to keep in mind."

All this depends on deploying lots of ICARUS tags, which cost about $500 for the first generation but could become cheaper and smaller in the next few years, says Wikelski. It will also require expanding the system from a lone antenna on the ISS to a network of satellites that would enable real-time read-outs and monitoring. He estimates that a three-satellite system would cost between $80 million and $100 million, and will require a lot of buy-in. "We have to show there's some value in global decision-making based on animal behaviour or movement," he says.

He also has to get ICARUS up and running. By mid-October, the Russians still had not switched ICARUS on for public use because of a snag in discussions between that nation and the DLR about the antenna's operation. Wikelski hopes those talks will be resolved soon and the system will come online. Then he needs biologists around the world to adopt it en masse — and soon. The ISS's Russian module is scheduled to operate for just six more years, although it could continue past that. The DLR, meanwhile, has plans to fund the mission only until 2024. "What comes after that is a big question," Weppler says.

That gives Wikelski and his collaborators a decade, at most, to convince the research community and space agencies that ICARUS is worth expanding into a global satellite network.

Days after his flight over Lake Constance, Wikelski was on a plane again — this time to Vancouver in Canada, where he announced the launch of ICARUS to the World Ornithological Congress. Over the next year, Wikelski will be travelling the globe to get ICARUS off the ground, helping to tag bears in Kamchatka in eastern Russia, condors in Bhutan, flying foxes in Zambia and migratory birds in the Congo Basin. "We have to go global," he says insistently. "We have to go wild. We have to go." ■

Andrew Curry *is a journalist in Berlin.*

1. Cochran, W. W., Mouritsen, H. & Wikelski, M. *Science* **304**, 405–408 (2004).
2. ICARUS. *White Paper: Global Satellite Tracking of (Small) Animals Will Revolutionize Insight Into Migration, Human Health, and Conservation* (ICARUS, 2008).
3. Sherub, S., Bohrer, G., Wikelski, M. & Weinzierl, R. *Biol. Lett.* **12**, 20160432 (2016).
4. Weinzierl, R. *et al. Ecol. Evol.* **6**, 8706–8718 (2016).

Cannabis samples grow under specialized lights at Anandia Labs in Vancouver.

JAMES MACDONALD FOR *NATURE*

A GOLD RUSH FOR CANNABIS

As legal weed hits Canada, scientists race to study and improve a once-forbidden plant.

BY ELIE DOLGIN

Jonathan Page has been around cannabis all his life. Growing up on Canada's Vancouver Island in the 1970s, he was surrounded by hippie beachcombers and dope smokers. So after earning a PhD in plant biology and phytochemistry, he felt completely at ease working with the plant *Cannabis sativa* as a postdoc in Germany in the early 2000s.

During that time, Page helped to characterize a pair of genes that some varieties of the plant uses to make fragrant oils responsible for pine- and lemon-like aromas [1]. And during an interview for a position with Canada's National Research Council (NRC), Page proposed similar projects to reveal how cannabis produces pharmaceutically active compounds known as cannabinoids.

He got the job, but was dismayed when he showed up to start his lab group in 2003 at the NRC's Plant Biotechnology Institute in Saskatoon, Saskatchewan. Page recalls his boss saying: "You're not going to work on cannabis here. We're the government."

What a difference a change in policy makes. On 17 October, Canada became the second country in the world, after Uruguay, to legalize cannabis for all uses. And although a few other countries, most notably Israel, have made a concerted effort to support agricultural research into cannabis, full legalization in Canada has brought with it unparalleled access to money for basic research on the plant.

Most of the country's 129 licensed cannabis producers are now clamouring to work with scientists on everything from gene mapping and metabolic engineering to optimal drying techniques and growing practices. And as part of an effort to corner the global legal cannabis market — one that's conservatively forecast to top US$57 billion within a decade — federal and provincial governments in Canada are putting up millions of dollars to support research.

Some researchers, such as Page (who still dabbled in cannabis research during his decade at the NRC), are well prepared to take advantage of Canada's great green rush. But botanists of all stripes are now turning to the plant, for both the funding opportunities and the uncharted science.

"You're talking about a plant that's a century out of date in terms of modern breeding techniques and scientific development," says Ernest Small, a botanist with Agriculture and Agri-Food Canada in Ottawa who has studied cannabis off and on since 1971.

RESEARCH BLUNTED

When Page first moved back to Canada 15 years ago, he initially resigned himself to studying a close relative of cannabis, the hop plant *Humulus lupulus,* which is used in brewing beer. But he doggedly pursued avenues to keep working on pot. Page secured a licence to grow industrial hemp, a variety of cannabis cultivated for its fibre that produces only trace amounts of tetrahydrocannabinol (THC), the mind-altering chemical responsible for cannabis's high. Eventually, he hooked up with the sole company contracted at the time by the government to

JAMES MACDONALD FOR *NATURE*

Many basic agricultural experiments have yet to be performed with cannabis.

produce the plant for medical purposes.

Page unpicked the pathway that leads to the formation of THC and cannabidiol (CBD) — cannabis's other main medically important compound[2]. Together with molecular geneticist Timothy Hughes at the University of Toronto, he sequenced the genome of a potent pot variety called Purple Kush[3]. But, says Page, "the NRC was still totally unsupportive of this work". So, in late 2013, he moved to Vancouver to start a cannabis biotechnology company called Anandia Labs.

In one of Anandia's first projects, Page worked with Sean Myles, a population geneticist at Dalhousie University's Agricultural Campus in Truro, Canada, to genetically characterize 124 samples of cannabis[4]. The analysis showed that the commercial labelling of subtypes indica and sativa rarely matched the plants' DNA profiles. And different samples marketed under the same madcap varietal name — White Widow, for example — often turned out to have wildly divergent genetics. "This is absolutely unthinkable in any other legitimate agricultural crop," says Myles. "You can't put a Mackintosh apple on the shelf and pretend it's a Honeycrisp."

Despite garnering publicity for the research, Page had trouble pulling in capital. Then came the election in October 2015 of Prime Minister Justin Trudeau, who promised during his campaign to legalize cannabis. "It changed attitudes almost overnight," Page says.

Although the Natural Sciences and Engineering Research Council of Canada doesn't have a dedicated cannabis initiative, the agency has funded dozens of projects focused on cannabis biology and cultivation. Genome Canada and other government-backed organizations have made research funds available, too. More substantially, private investment dollars have come pouring into the country's cannabis industry (see 'A smoking-hot sector'). Last year alone, Canadian cannabis companies raked in close to Can$2 billion (US$1.5 billion) — more than half of all the funding raised by legal cannabis firms worldwide — and the industry is on track to triple that number in 2018.

Anandia was one of the many beneficiaries. After securing more than Can$13 million (US$10 million) in private investment, the company got snapped up earlier this year by industry heavyweight Aurora Cannabis in Edmonton, for Can$115 million. "That is a major gesture of confidence," says Cam Battley, chief corporate officer at Aurora, adding that science and innovation are key to growing "a globally competitive company that will be built to last".

The sentiment is a relatively new one, says Michael Ravensdale, a plant pathologist who leads production at the firm CannTrust in Vaughan. "Science was in short supply, but it's going to be very important for the next chapter of the cannabis industry."

That's why many companies investing in research are starting with the fundamentals. "There are these super basic, huge questions that need to be answered," says Greg Baute, who used to breed tomatoes at Monsanto in Woodland, California, and moved north this year to head breeding and genetics at Anandia's new research facility in Page's home town of Comox. "You can do these really straightforward experiments and get these huge results."

JOINT VENTURE

In a suburban Toronto mall, nestled alongside a paint store, sits a nondescript brick building, home to TerrAscend, a cannabis producer that has been shipping its product for about a month in anticipation of 17 October. Inside, past a barbed-wire fence and several layers of electronic security, are grow rooms full of Shishkaberry, CBD God Bud and Cold Creek Kush, cultivars valued for their sleep-inducing, antidepressant and stress-relieving properties, respectively.

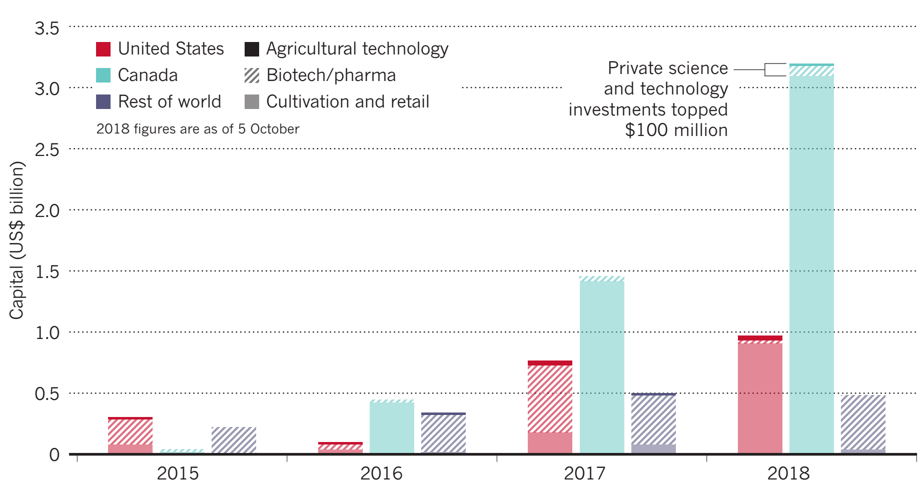
Nearly all the plants are unpollinated females, called 'sinsemilla' (meaning 'without seed'), that produce the highly potent flowering tops, or buds, rich in THC and other cannabinoids. Males, with their pollen-filled sacs, are not only superfluous — the female plants are all propagated through cuttings — but also avoided for fear of scrambling genes in an uncontrolled way.

Yet, TerrAscend has a small space in the back reserved for males. Back in June, the company launched a research and development arm in conjunction with scientists at two Ontario universities who will use the sequestered plants for experiments that are standard in agriculture but have rarely been done with cannabis. The scientists will mate the plants, coax them to produce seeds and then expose the seeds to chemical mutagens in the hope of finding new desirable traits — pest resistance, say, or increased tolerance to environmental stresses such as drought.

Scientists involved with the TerrAscend

A SMOKING-HOT SECTOR

Business deals in the cannabis industry have spiked in Canada since the nation decided to legalize recreational use. But most of the capital has gone to companies involved in growing and selling. Canadian companies' investments in research and development are catching up to the levels seen elsewhere in the world.



SOURCE: VIRIDIAN CAPITAL ADVISORS

THE WIDE WORLD OF WEED

Legal restrictions haven't completely stymied cannabis research.

Although cannabis is now legal in some form in all US states but one, the plant, including hemp, is still illegal at the federal level — and all research conducted on cannabis at any university must abide by federal regulations, or else jeopardize government funds for the institution.

Extracted DNA samples are permitted, and the government is gradually becoming more permissive of hemp cultivation. But any work on the basic biology of cannabis with higher levels of the psychoactive substance tetrahydrocannabinol (THC) is strictly off-limits, and there's no funding available for non-health-related research on the plant. "Everything we've done has been absolutely shoestring," says plant biologist George Weiblen at the University of Minnesota in Saint Paul.

That's led to a few creative workarounds. In Colorado, where cannabis has been legal for recreational use since 2014, evolutionary geneticist Daniela Vergara, at the University of Colorado Boulder, created a foundation to support projects she can't do on campus. Unlike her university, the foundation can accept funding from the cannabis industry — money she then spends on sampling DNA from cannabis plants bred at local dispensaries and growers. Although she can analyse DNA sequence data at the university, "I don't touch the plant on campus or during my working hours," Vergara says.

Even at the University of Mississippi in Oxford, which is licensed by the federal government to grow cannabis for health studies, researchers are operating under restrictive rules. The university can procure plants only from federally approved vendors, and so it has no access to the THC-rich varieties commonly found in recreational and medicinal cannabis shops. Instead, scientists there are using chemical stimulants to boost THC levels in moderate-strength plants. "My hands are tied," says Mahmoud ElSohly, who oversees the operation. "I have to work with what I have."

Even in Uruguay where cannabis is completely legal, hurdles still exist. In 2013, the country became the first to legalize cannabis for recreational use, but researchers have faced bottlenecks in gaining access to the plant, according to an analysis[12] in March, and there's no public money to foster research. "I get very, very little funding," says Astrid Agorio, a plant molecular geneticist at the Clemente Estable Institute of Biological Research in Montevideo who is attempting to exhaustively profile the genetic structure of two varieties of Uruguayan cannabis with a grant of about US$6,000.

In Australia, where medical use is allowed, plant geneticist Graham King at Southern Cross University in Lismore has found both public and private support. He obtained a licence from the state of New South Wales to grow cannabis, and financing from an industrial hemp company to chemically characterize a global collection of cannabis varieties[13]. "We want to understand what the scope is for metabolic engineering," King says.

But nowhere has been as supportive of cannabis research as Israel. It was here that Raphael Mechoulam, an organic chemist, isolated THC and cannabidiol and determined their chemical structures in the 1960s. The country has since sought to establish itself a world leader in the study of medical cannabis. And unlike in some countries, including Canada, where government scientists still essentially can't access high-THC varieties of the plant, Israel's agriculture ministry last year built a national centre for medical-cannabis research.

Housed at the Volcani Center in Rishon LeZiyyon, the government's cannabis farm includes thousands of plants spread across several greenhouses and indoor growing facilities. There, plant biologist Nirit Bernstein is trying to perfect cultivation and weed out bad practices. "We have to develop science-based protocols for optimizing the cultivation of this magical plant," Bernstein says. "There's very little scientific information that's available." **E.D.**

spin-off, including plant geneticists Peter McCourt and Shelley Lumba at the University of Toronto, plan to mutagenize six varieties of cannabis with the aim of obtaining improved versions of some of the company's go-to stock. "Our main goal," says Lumba, "is to make cannabis into a real horticultural crop."

Another decades-old practice for improving agricultural plants involves intentionally doubling or tripling their genomes, which tends to give plants bigger cells, larger structural features and greater yields of chemical compounds. Domesticated wheat species, for example, have 4–6 copies of their genome; sugar cane can have as many as 16. And although most modern farmed plants had their DNA multiplied simply through hundreds of years of cultivation, there are ways to speed up the process. All cannabis varieties characterized so far have only two sets of the genome — all, that is, except for a handful of plants growing at Canopy Growth Corporation in Smiths Falls.

There, plant molecular geneticist Shelley Hepworth at Carleton University in Ottawa and her former graduate student used a cell-cycle-disrupting herbicide to trigger five varieties of cannabis to double their normal chromosome count. At first blush, says Hepworth, "the plants are definitely bigger". But the scientists still need to finish their analyses to determine whether the 'tetraploid' cannabis lines have elevated levels of THC, CBD or other cannabinoids.

A more modern plant-breeding technique — one that dates back to the 1980s — is known as marker-assisted selection. It involves finding genetic signatures associated with a desirable trait — high essential-oil content, say, or automatic flowering under any light condition. Scientists can then use DNA analyses to quickly 'preview' which seedlings should have optimal properties instead of waiting months for the plants to mature.

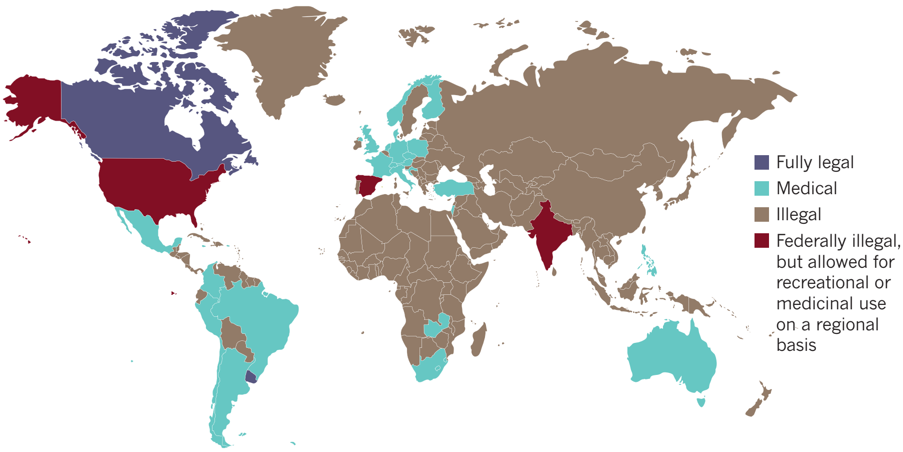
"THIS YEAR IS A TURNING POINT FOR CANNABIS MOLECULAR GENETICS."

Only a small number of markers exist for cannabis, however, in large part because few researchers have ever looked for them. One that has been described comes from George Weiblen, a plant biologist at the University of Minnesota in Saint Paul and one of the few academics in the United States who has a federal licence to grow cannabis — but he's restricted to just 50 plants at a time (see 'The wide world of weed'). It took him 12 years to determine the inheritance pattern of genes that affect drug content and to identify a genetic marker linked to the THC-to-CBD ratio[5]. That's longer than it took Gregor Mendel to work out the laws of heredity, Weiblen says. "Our programme is a poster child for the absurdity of cannabis research in the United States."

A complete cannabis genome would make it easier to identify informative DNA markers. But early efforts have yielded maps that are patchy and incomplete, says Kevin McKernan, chief science officer and founder of Medicinal Genomics in Woburn, Massachusetts, who did some of the early work. He (see go.nature.com/2kpzgkc) and Page[3] released maps

A BUDDING TREND

Political movements to decriminalize cannabis for medical and recreational use have been gaining in popularity. The result is a slowly shifting patchwork of regulations as many countries (and sometimes smaller jurisdictions) start to allow limited use and increasingly liberalize their stance.



SOURCE: ARCVIEW MARKET RESEARCH/BDS ANALYTICS

independently in 2011. "They're train wrecks," says McKernan. But that's changing. And thanks to an international trend toward less restrictive laws around cannabis (see 'A budding trend'), many weed companies are now investing in genetics research.

DANK DNA

"This year is a turning point for cannabis molecular genetics," says C. J. Schwartz, founder and chief executive of Sunrise Genetics in Fort Collins, Colorado — one of at least six companies that say they have come up with fine-scaled genome maps. They have not yet been published yet in a peer-reviewed journal, but McKernan posted a preprint of his map on 10 October[6] and Schwartz expects to make his sequence public by the end of the month.

Then there are scientists who are hoping to engineer new properties into the plant. At Canopy Growth, research and development manager Katya Boudko worked with molecular biologist Douglas Johnson at the University of Ottawa to develop a gene-silencing technology that prevents expression of the THC-synthesis gene. Boudko expects plants to compensate by boosting their levels of CBD — or, she says, "it could potentially produce other cannabinoids that the world doesn't even know about yet".

Boudko has yet to fully test this theory, however. That's because she hasn't managed to grow fully fledged plants from genetically modified tissue — and nor have many others. Because seeds or clippings cannot be genetically modified in a consistent and predictable way, scientists need to culture plant tissue and coax it into producing roots and shoots after the genes have been manipulated. Often, scientists can get the cellular masses to produce fine root hairs, but the shoots have proved particularly problematic.

In 2010, a team from the University of Mississippi in Oxford, where for 50 years researchers have grown all the cannabis used for government-backed health studies in the United States, described a hormone recipe for inducing shoot formation that, it says, works more than 80% of the time[7]. Yet, others say that they can't get the protocol to work on their own varieties. "There are tens of labs that are now working on making the protocol more efficient," says Leor Eshed-Williams, a plant developmental geneticist at the Hebrew University of Jerusalem.

For many in the industry, however, any suggestion of genetic modification is an anathema. Besides, says Ethan Russo, director of research and development at the International Cannabis and Cannabinoids Institute in Prague, "this plant is so malleable that a lot of these genetic modifications are really unnecessary". Modern selective-breeding strategies, he argues, should suffice — and those methods needn't require

"THE PLANTS, AT THE END OF THE DAY, NEED LOVE."

genetic markers. In collaboration with Mark Lewis, president of Napro Research in Westlake Village, California, Russo has used chemical profiling to create dozens of cannabis varieties with unique properties and elevated yields[8].

HASHING OUT THE BASICS

Elsewhere, researchers are looking to control and fine-tune environmental conditions at various stages of the growth cycle. Such tweaks could prove important for maximizing profits from costly indoor growing operations that are a major source of high-end cannabis in Canada. At CannTx Life Sciences in Puslinch, operations head Jeff Scanlon and his colleagues developed a system for air circulation.

Scanlon showed that the fans found in most companies' grow rooms move air over the plant crowns. But in the thicket of leaves and branches beneath, the air remains stagnant, leading to pockets of elevated temperature and humidity that breed fungal pathogens. The solution: a pressure gradient from floor to ceiling that ensures airflow along every surface of the plant. "It's a very simple innovation," Scanlon says.

Deron Caplan, director of plant science at Flowr in Lake Country, completed a PhD this year in which he systematically determined optimal rates of fertilizer supply at various stages of cannabis production[9,10] and best practices for propagating the plant clonally through cuttings[11]. "It's very crude, incremental advancements that we're making," says Mike Dixon, an agricultural scientist at the University of Guelph and one of Caplan's graduate advisers. But it's slowly helping to phase out the homespun practices that persist throughout much of the industry, he says — "what I kindly refer to as anecdotal bullshit".

Some of the holdovers of illegal cultivation are on display at Beleave, a cannabis producer in Hamilton, where master grower Shane Whelan-Stubbs has persevered with some practices for 20 years, first in a bedroom cupboard, then in basements, warehouses and now a legitimate business. Whelan-Stubbs is open to the science, and Beleave will soon start collaborating with a team at Guelph, where scientists hope to open the country's first dedicated academic centre for cannabis research some time next year. Still, Whelan-Stubbs continues to water his plants by hand. "The plants, at the end of the day, need love," he says.

Page would argue that they also need science. But the burgeoning research that he's been a part of likewise would have long ago withered without cannabis at its centre. "We think of it in many ways as a drug or a pharmaceutical," he says, "but can't forget that it's the plant that's at the heart of this revolution. It all comes down to a plant." ■

Elie Dolgin *is a Canadian-born science journalist in Somerville, Massachusetts.*

1. Günnewich, N. *et al. Nat. Prod. Commun.* **2**, 223–232 (2007).
2. Gagne, S. J. *et al. Proc. Natl Acad. Sci. USA* **109**, 12811–12816 (2012).
3. van Bakel, H. *et al. Genome Biol.* **12**, R102 (2011).
4. Sawler, J. *et al. PLoS ONE* **10**, e0133292 (2015).
5. Weiblen, G. D. *et al. New Phytol.* **208**, 1241–1250 (2015).
6. McKernan, K. *et al. OSF Preprints* https://doi.org/10.31219/osf.io/7d968 (2018).
7. Lata, H., Chandra, S., Khan, I. A. & Elsohly, M. A. *Planta Med.* **76**, 1629–1633 (2010).
8. Lewis, M. A., Russo, E. B. & Smith, K. M. *Planta Med.* **84**, 225–233 (2018).
9. Caplan, D., Dixon, M. & Zheng, Y. *HortScience* **52**, 1307–1312 (2017).
10. Caplan, D., Dixon, M. & Zheng, Y. *HortScience* **52**, 1796–1803 (2017).
11. Caplan, D., Stemeroff, J., Dixon, M. & Zheng, Y. *Can. J. Plant Sci.* **98**, 1126–1132 (2018).
12. Hudak, J., Ramsey, G. & Walsh, J. *Uruguay's Cannabis Law: Pioneering a New Paradigm* (The Brookings Institution, 2018).
13. Welling, M. T., Liu, L., Shapter, T., Raymond, C. A. & King, G. J. *Euphytica* **208**, 463–447 (2016).

CHRIS HONDROS/GETTY

A UN peacekeeper at an election in Liberia in 2005.

Retool AI to forecast and limit wars

Using artificial intelligence to predict outbursts of violence and probe their causes could save lives, argue **Weisi Guo**, **Kristian Gleditsch** and **Alan Wilson**.

Armed violence is on the rise and we don't know how to stop it[1]. Since 2011, conflicts worldwide have killed up to 100,000 people a year, three-quarters of whom were in Afghanistan, Iraq and Syria. The rate of major wars has decreased over the past few decades. But the number of civil conflicts has doubled since the 1960s, and terrorist attacks have become more frequent in the past ten years.

The nature of conflict is changing. Wars are waged less often between states, but increasingly within them by armed groups — more than 1,000 such groups operated in Syria at the peak of its civil war in 2013. They vary in size from a few local militias to tens of thousands of experienced fighters. Advances in technology makes attacks more precise, coordinated and deadly. Civilians are increasingly targeted. By 2016, wars had displaced more than 65 million people worldwide from their homes. More than half were children.

The costs are huge. The United Nations spent more than US$20 billion in 2016 on humanitarian aid. Violent countries are weakened economically. For example, since 1996, wars have cost the Democratic Republic of the Congo almost one-third of its gross domestic product[2]. Wars stifle progress towards many of the UN Sustainable Development Goals.

Nations spend relatively little on preventing conflicts. UN peacekeeping efforts in 2016–17 cost around $7 billion, equivalent to less than 1% of global military spending. Yet peacekeepers have prevented conflicts from erupting in the wake of crises[3]. For example, within one month of a disputed presidential election in Gambia in 2016, West African countries sent troops to maintain security. And interventions can stop them from ▶

CONFLICT PREDICTION

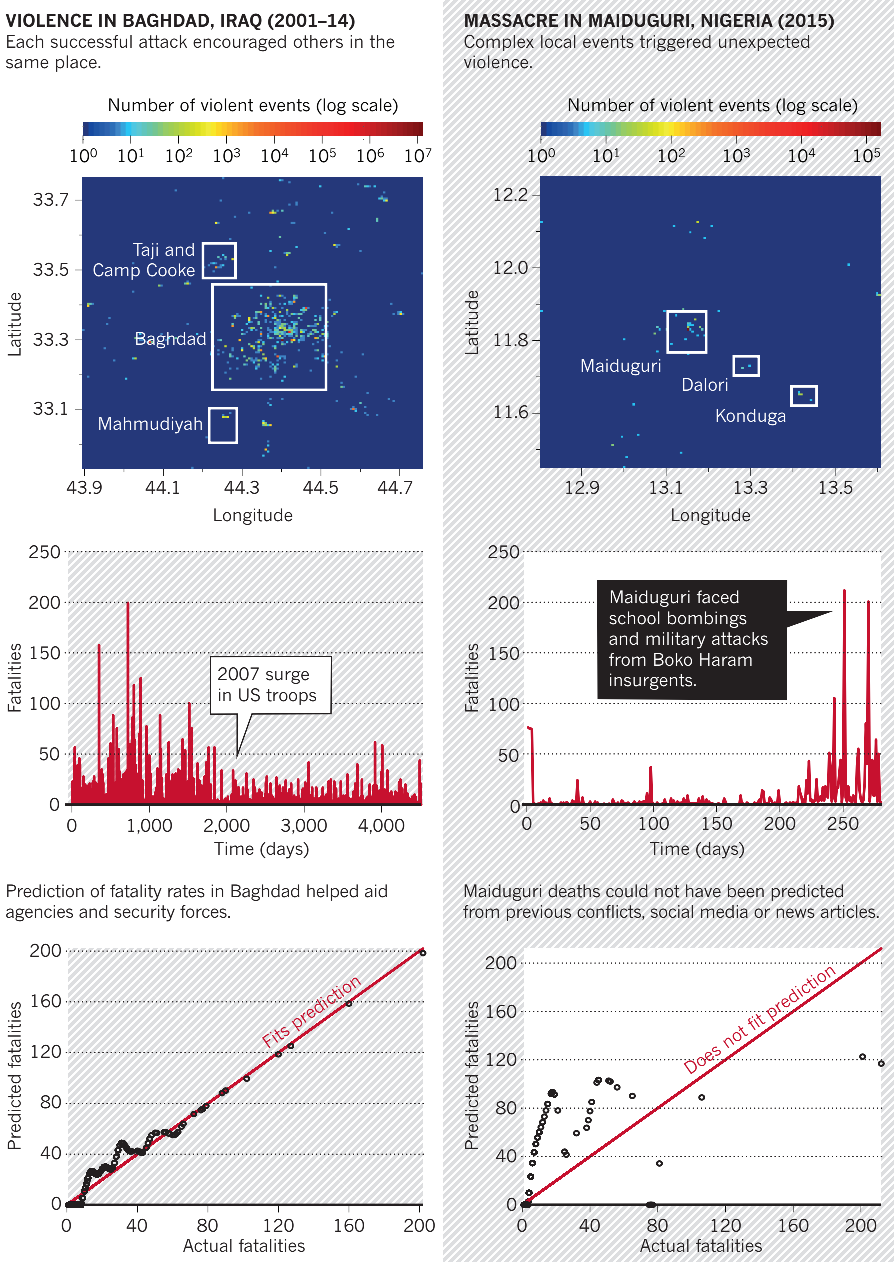
Armed conflicts differ in the degree to which they can be forecast.

WHITE SWAN EVENTS

Escalate on their own, follow statistical laws and can be predicted.

BLACK SWAN EVENTS

Erupt without precedent, don't follow statistical laws and cannot be predicted.



SOURCE: GLOBAL TERRORISM DATABASE, HTTPS://WWW.START.UMD.EDU/GTD

recurring, as in El Salvador's civil war in 1991 and in Bosnia and Herzegovina in 1995.

Governments and the international community often have little warning of impending crises. Likely trouble spots can be flagged a few days or sometimes weeks in advance using algorithms that forecast risks, similar to those used for predicting policing needs and extreme weather. For conflict risk prediction, these codes estimate the likelihood of violence by extrapolating from statistical data[4] and analysing text in news reports to detect tensions and military developments (see go.nature.com/2oczqep). Artificial intelligence (AI) is poised to boost the power of these approaches.

Several examples are under way. These include Lockheed Martin's Integrated Crisis Early Warning System, the Alan Turing Institute's project on global urban analytics for resilient defence (run by W.G. and A.W.) and the US government's Political Instability Task Force.

Future AI and conflict models need to do more than make predictions: they must offer explanations for violence and strategies for preventing it. This will be difficult because conflict is dynamic and multi-dimensional. And the data collected today are too narrow, sparse and disparate.

Three things will improve conflict forecasting. They are: new machine-learning techniques; more information about the wider causes of conflicts and their resolution; and theoretical models that better reflect the complexity of social interactions and human decision-making.

BROADEN DATA COLLECTION

Conflict prediction took off in the 1920s and 1930s. One of its pioneers was mathematician Lewis Fry Richardson, who applied statistics to study the causes of wars[5]. He revealed certain rules, such as that there are more small fights that kill a few people than large ones in which many die. Gang murders in Chicago in Illinois and Shanghai, China, followed the same scaling laws as major wars, he found. These laws tell us roughly how many skirmishes to expect, but not where or when they might occur.

Data collection has marched on. Fatalities, locations, actors and objectives for hundreds of thousands of battles and attacks are recorded in databases such as the Armed Conflict Location and Event Data Project (www.acleddata.com), the Global Terrorism Database (www.start.umd.edu/gtd) and the Uppsala Conflict Data Program (http://ucdp.uu.se). The data come from many sources, often media reports, and are checked by human specialists. They are good enough to give authorities a few days' notice of worse to come, but not the weeks or months of warning that are needed to devise strategies for peaceful resolutions.

The types of data collected, and the predictive models, are too crude to reveal the social drivers of conflict. Sometimes the most important outcomes are unobservable: incidents that were deterred or thwarted by security forces or back-door political bargaining. Media reporting of violence is stifled in countries such as Iran. Actors can shift tactics and allegiances. Violent factions might stoke tensions in the background while pursuing peace in public, as in Northern Ireland during the Troubles (1966–1998) and in Colombia since 1964.

Levels of violence depend on intangibles such as the willingness to fight. Weapons and funds from sources outside the country intensify civil conflicts, as in Syria and Yemen. Successful attacks encourage further attempts. Inequality, ethnic tensions or oppressive governance can trigger riots or civil wars. Environmental factors such as drought add to all these pressures.

REDUCE UNKNOWNS

AI will add little if social and causal factors are omitted. Furthermore, the statistical approaches used today for machine learning cannot deal with such a mix of unknown information. For instance, AIs need to be

LUIS ROBAYO/AFP/GETTY

Members of the FARC guerrilla group in Colombia began a disarmament process in 2017.

trained to make inferences. They 'learn' from existing data, test whether predictions hold, and then hone the algorithms accordingly. This assumes that the training data mirror the situation being modelled. The problem is that, often, we do not know whether they are similar or not, especially in evolving scenarios that involve many hidden factors. If they don't, then the predictions are unreliable.

More needs to be learned about the statistics of different types of conflict; we already know that there are differences (see 'Conflict prediction'). For example, separatist ethnic conflicts are more likely to stay within a homeland than spread beyond it. Terrorist attacks are more common in civil wars than in invasions. Elaborate attacks requiring planning, such as the terrorist strikes on New York's World Trade Center on 11 September 2001, are most likely to occur outside a conventional conflict.

DEVELOP THEORIES

Conflict researchers have yet to develop a universally agreed framework of theories to describe the mechanisms that cause wars. Such a framework would dictate which sorts of data are collected and what needs to be forecast. Most current studies test data against simple informal hypotheses, such as that climate change increases violence. Correlations are sought; models disagree; the results are contentious[6]. Too few questions are asked about context, such as political and economic inequalities or military deterrence.

Modelling complexity will be key. For example, where is it best to intervene for a peaceful outcome, and how much intervention is needed? Algorithms could tease out spatial patterns from interacting stakeholders[7], or highlight unstable geographical boundaries using the theory of social competition between neighbouring groups[8]. For example, a team including one of us (A.W.) used such models[9] to characterize the London riots in 2011 after the event. The work confirmed the police numbers that were required to restore order. However, models such as these are only as good as the data that go into them.

A GLOBAL CONSORTIUM

Conflict prediction and prevention need a global data-driven system, like those for forecasting weather, epidemics and maintenance needs in engineering. We propose that an international consortium be set up to develop formal methods to model the steps society takes to wage war. Establishing this platform would cost tens of millions of dollars, a fraction of the billions that the world pays to cope with conflict.

The consortium should involve academic institutions, international and government bodies (such as the European Commission Disaster Risk Management Knowledge Centre, UN Peacekeeping and national foreign offices) and industrial and charity interests in reconstruction and aid work (such as the engineering and construction consultancy Arup, and the International Red Cross and Red Crescent Movement). Academic researchers should set up a virtual global platform for comparing AI conflict algorithms and socio-physical models[10]. This must use open-access data to accelerate reproducible research and to benchmark outputs. Standards for measurements, theories and models need to be developed.

We hope to take the first steps to agree a common data and modelling infrastructure at a workshop on 15–16 October. The event, organized by Uppsala University in Sweden, will focus on the Violence Early-Warning System (ViEWS; see go.nature.com/2y7b9qt). We call on the UN to invest in data-driven predictive methods for promoting peace. ■

Weisi Guo *is a Turing Fellow and associate professor in the School of Engineering, University of Warwick, Coventry, UK.* **Kristian Gleditsch** *is professor of political science in the Department of Government, University of Essex, Colchester, UK.* **Alan Wilson** *is director of special projects at the Alan Turing Institute, London, UK.*
e-mail: weisi.guo@warwick.ac.uk

1. World Bank & United Nations. *Pathways for Peace: Inclusive Approaches to Preventing Violent Conflict* (World Bank & United Nations, 2018).
2. IANSA, Oxfam & Saferworld. *Africa's Missing Billions* (IANSA, Oxfam & Saferworld, 2007).
3. Hegre, H., Hultman, L. & Nygard, H. *Peacekeeping Works: An Assessment of the Effectiveness of UN Peacekeeping Operations* (Peace Research Institute Oslo, 2015).
4. Zammit-Mangion, A., Dewar, M., Kadirkamanathan, V. & Sanguinetti, G. *Proc. Natl Acad. Sci. USA* **109**, 12414–12419 (2012).
5. Richardson, L. F. *Nature* **155**, 610 (1945).
6. Adams, C., Ide, T., Barnett, J. & Detges, A. *Nature Clim. Change* **8**, 200–203 (2018).
7. Turchin, P., Currie, T. E., Turner, E. A. L. & Gavrilets, S. *Proc. Natl Acad. Sci. USA* **110**, 16384–16389 (2013).
8. Lim, M., Metzler, R. & Bar-Yam, Y. *Science* **317**, 1540–1544 (2007).
9. Davies, T. P., Fry, H. M., Wilson, A. G. & Bishop, S. R. *Sci. Rep.* **3**, 1303 (2013).
10. Caldarelli, G., Wolf, S. & Moreno, Y. *Nature Phys.* **14**, 870 (2018).

NUTRITION

Poisoned platefuls

Felicity Lawrence extols two chronicles on the long, fierce fight for US food safety.

In 1902, the US Congress funded the first controlled trials of food toxicity involving human participants. The chief chemist of the US Department of Agriculture (USDA), Harvey Washington Wiley, was given US$5,000 to investigate how food preservatives and colourings affected health. It was a key moment in a long and ongoing fight to stop industry riding roughshod over the public interest in the supply of food.

Wiley recruited young, healthy men as guinea pigs, starting with civil servants. They signed liability waivers and agreed to take part in "hygienic table trials", eating free but strictly prescribed meals in an experimental kitchen in the USDA's basement in Washington DC. An excitable press dubbed them the Poison Squad, giving Pulitzer-prizewinning science journalist Deborah Blum the title for her meticulous book tracking the early history of US food regulation. Meanwhile, Marion Nestle, academic scourge of 'Big Food', brings the account up to date in *Unsavory Truth*, her latest withering analysis of industry efforts to corrupt science and dodge regulation.

The Poison Squad: One Chemist's Single-Minded Crusade for Food Safety at the Turn of the Twentieth Century
DEBORAH BLUM
Penguin Press (2018)

Unsavory Truth: How Food Companies Skew the Science of What We Eat
MARION NESTLE
Basic (2018)

As Blum's chronicle reveals, two rapidly developing industries untrammelled by government oversight came together to disastrous effect. The second half of the nineteenth century had seen an explosion in US chemical manufacturing as the country shifted from an agricultural economy to an increasingly industrialized and urbanized one. Newly synthesized preservatives were cheap, and were added liberally to all sorts of food. Refrigeration was still in its infancy, and not yet adapted for domestic use.

Meat, tinned fruit and vegetables, butter and cheese were dosed with boric acid, salicylic acid and sodium benzoate to delay bacterial growth and rotting. Commercial butchers found that salicylic acid set off a chemical reaction that made old, greying meat look freshly pink for 12 hours. Formaldehyde, the embalmer's tool, was a favourite for treating milk about to go off: its sweet taste masked rancidity. Along with newly developed coal-tar dyes and other toxic stalwarts of food colouring, such as chromate of lead (used to turn sweets yellow), these chemicals were deployed to disguise adulteration and dangerous spoilage. There were many mass poisonings; in 1899, 400 children in Indiana died after drinking "embalmed" milk. Yet there were no federal laws at the time covering the sale of unsafe food, nor

ILLUSTRATIONS BY THOMAS PATERSON

any requiring accurate labelling.

From his appointment in 1883, Wiley had run numerous tests exposing this widespread adulteration of food and drink, infuriating powerful interests — from dairy and meat producers to whiskey distillers. By 1902, he was battle-hardened and adept at working with writers and women's groups to promote national regulation of the food sector.

His Poison Squad experiments proved decisive. The first group of 12 volunteers was divided into two. Half had their food dosed at varying levels with chemical preservatives, starting with borax, a salt of boric acid; the others ate the same meals, free of additives. The groups were then swapped. Temperatures and pulses were recorded and monitored, urine and faeces collected and analysed. Double blind it most certainly was not. The participants soon worked out that the borax had been secreted in butter, and stopped putting it on their bread. The USDA scientists finally resorted to administering the preservative direct to the volunteers, in capsules.

The project's chef had a loose tongue, and the press ran lurid stories, exposing government chemists to ridicule. Wiley persevered. His guinea pigs became ill: effects ranged from confusion to nausea and vomiting, increasing with cumulative dosing. The case for legislation was becoming irrefutable, and despite the efforts of industry allies in both the House of Representatives and the Senate, US President Theodore Roosevelt was coming round to supporting it. (Lobbying and donations from the food-manufacturing, whiskey and chemicals industries to politicians and scientists were as liberally dispensed as the preservatives.) When the Pure Food and Drug Act passed in 1906, it became widely known as "Dr Wiley's law".

Changing government policy is rarely fast; instead, it is a series of protracted skirmishes and incremental changes. Blum's chronological narrative in *The Poison Squad* sometimes gets bogged down in minutiae, much as campaigners did. One chapter also threatens to run away with the book. No history of US food regulation would be complete without Upton Sinclair, the young socialist "muckraking" writer who documented the horrors of the stockyards in Chicago, Illinois. Blum's narrative on Sinclair's 1906 novel about it, *The Jungle*, risks upstaging her hero. Sinclair based his book on seven weeks observing the brutal conditions, as immigrants worked with diseased cattle and a hellish mix of rotting meat, floor sweepings, carcasses retrieved from privies, rats and rat poison, which were all processed together. That proved the tipping point for Roosevelt to back legislation. Blum's account of Wiley's work is full of fascinating detail and is a valuable contribution to understanding the politics of food.

Nestle, a nutrition researcher at New York University, writer and distinguished veteran of many an advisory committee, could make a fair claim to Wiley's mantle today. For decades, she has been battling the food and drink industry, with a combination of sound science and brilliant communication. Like Wiley, she has found herself becoming part of the story, attacked in the media for exposing adulterations and routine poisonings — albeit a less acute and more chronic epidemic of them than Wiley's, in the form of diet-related non-communicable diseases, such as obesity, diabetes and cardiovascular disease. Her earlier works, notably *Food Politics* (2002) and *Safe Food* the following year, were key examinations of the problems of today's food supply. She must sometimes long for the simplicity of the Poison Squad experiments: a theme of *Unsavory Truth* is the complexity of nutrition research. The book is a remorseless dissection of the corruption of science by industry.

The food industry's playbook is familiar from the strategies of tobacco and climate-change denial over the past four decades. Yet it is poorly understood, and ignored by some media and academic journals in the field. It relies, as Nestle points out, on repeated use of the same set of techniques. Cast doubt on unhelpful science; fund more favourable, skewed science; offer gifts and consultancies; sponsor professional bodies; and use front groups posing as independent institutes. Finally, promote personal responsibility and self regulation rather than government intervention; capture advisory committees; and challenge regulation in court.

Too many industry-funded studies posing as serious scientific inquiry are in fact marketing research for single products or ingredients. She demolishes claims from a chocolate-milk drink that purportedly helps young American footballers' cognitive function even after concussion, to blueberries touted as preventing erectile dysfunction. She asks why serious journals publish these. Her extensive review shows the vast majority of such studies are favourable to the funder, whereas in independent research the opposite is the case. Yet researchers in nutrition, as in other fields, are under intense pressure to bring in grants whatever the conflicts of interest. Nestle is more generous than I might be in exonerating many from conscious bias, arguing that the studies themselves are often good science but the problem lies in who frames the questions and how the results are interpreted.

Nestle's accounts of conflicts of interest include Coca-Cola funding university researchers in a "Global Energy Balance Network", focusing obesity studies on physical activity rather than diet. Reading these, it's hard to argue with her call for full disclosure, and recognition and active management of those conflicts. But the answers are much bigger than that, as she acknowledges. "Corporations have taken over American society, putting democratic processes at grave risk," she notes.

TOO MANY STUDIES POSING AS SERIOUS SCIENTIFIC INQUIRY ARE IN FACT MARKETING RESEARCH.

Her solution? Engaged citizens and better rules that "control the political power corporations exert over legislation and policy". These two books about the troubled history of food safety demonstrate that science does not sit in a protected space of apolitical empirical truth. Like everything else, it is part of the battleground that is politics. ■

Felicity Lawrence *is author of* Not on the Label, *and is an Orwell-prizewinning writer for* The Guardian *in London.*
e-mail: felicity.lawrence@theguardian.com

HISTORY

Reimagining the dog

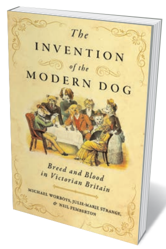
Meg Daley Olmert enjoys a story of the Victorian drive towards unnatural selection.

Charles Darwin, Charles Dickens and P. T. Barnum walk into a pub … a classic comic set-up that can only lead to one punch line: *The Invention of the Modern Dog*. This chronicle — by science historians Michael Worboys and Neil Pemberton and historian Julie-Marie Strange — charts the confluence of biology, class and popular entertainment that resulted in an unprecedented burst of nineteenth-century canine breeding. That tumult, they argue, stares out at us today from the eyes of our dogs.

Science and engineering were reshaping the British Isles, shrinking distances both geographical and social, even as Darwinian science effectively 'shrank' the distance between species. A newly minted middle class donned morning coat and top hat, and strode off to support the making of Empire. Vast numbers of people were 'improved' as their hard work finally paid off, and improved people needed improved dogs. Between 1874 and the beginning of the twentieth century, the number of dog breeds recognized in Britain swelled to include foreign breeds, variations of older ones such as the Welsh spaniel and the Skye terrier, and "manufactures", such as the Yorkshire terrier. These dogs were whipped into must-have status through another Victorian invention: the dog show.

For the lucky and industrious, there was much to celebrate, and the money and time to do it. Over more than 270 pages, the authors document the dog show "mania" that swept across Britain from 1862. It is here that the new, improved dog took centre stage, for better or worse.

The British aristocracy had always been keen on dogs: stud pedigrees remain as closely tracked and controlled as those of their masters. Canine valour was tested in the field and exalted in the arts. Meanwhile, the dog as entertainment had long been the domain of the 'lower' classes. Bull and bear baiting, popular in Elizabethan London, ended only in 1835, with the first Cruelty to Animals Act. Rat baiting remained a gambling sport in pubs until 1912. Aficionados of these 'entertainments' were known as "the Fancy". Those who bred the fastest, toughest dogs could make a good living on wagers set by the thirsty new army of clerks and mid-level managers. For their wives, the new money bought servants, and empty time and laps were soon filled by miniature dogs such as the King Charles spaniel favoured by Queen Victoria (see *Nature* http://doi.org/gdthxg; 2018).



The Invention of the Modern Dog: Breed and Blood in Victorian Britain
MICHAEL WORBOYS, JULIE-MARIE STRANGE & NEIL PEMBERTON
Johns Hopkins University Press (2018)

The Fancy scaled up its 'sporting' events and toy-dog beauty shows. The gentry, seeing this as entrepreneurial overstep threatening to dilute the purity of the British dogs' pedigree, created the Kennel Club to set rules for the shows. Yet, despite its claimed dedication to improving breeds, the club never set breeding standards. Those were the preserve of local clubs devoted to a single breed, a newly emerged social stratum spanning the amateur–professional chasm.

Enter Darwin, Dickens and Barnum. Unnatural selection, social pretentions and showbiz set the tone for thousands of dog shows drawing Victorians of all classes (on separate days) to marvel at dogs that were changing before their eyes. With wolves extinct in Britain, animal baiting banned and game birds bred and delivered within shooting range, dogs no longer needed valour, courage and stamina. Freed to select for conformation alone, each club created an exacting standard for its breed's appearance and assigned a numerical value to it.

There was little science to guide them. They did have Darwin's warnings about the evils of inbreeding; and a Lamarckian belief in the heritability of acquired traits still lingered. The well-established practice of outbreeding periodically to improve performance was cast aside in favour of inbreeding to produce physical duplicates of the latest standard. 'Best of show' would go only to a black Newfoundland. The pug was shrunk from 14 kilograms to 10. Pointers grew larger. More toxic standards were set for the newly redundant bulldog. Selective breeding and surgical 'face jobs' produced extremely flat-faced dogs that were favoured in the show ring but were reportedly devoid of courage and aggression. George Roper, a leading figure in the Bulldog Club, lamented that the breed was "more liable to deterioration" than others.

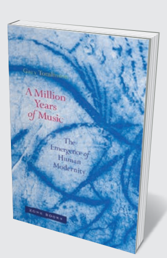
THE GOAL WAS IMPROVEMENT OF THE DOG SHOW, NOT THE DOG.

The goal was improvement of the dog show, not the dog. Breed standards, based on fashion, were locked into place to make judging easier and competition fairer. The result was dog-as-commodity.

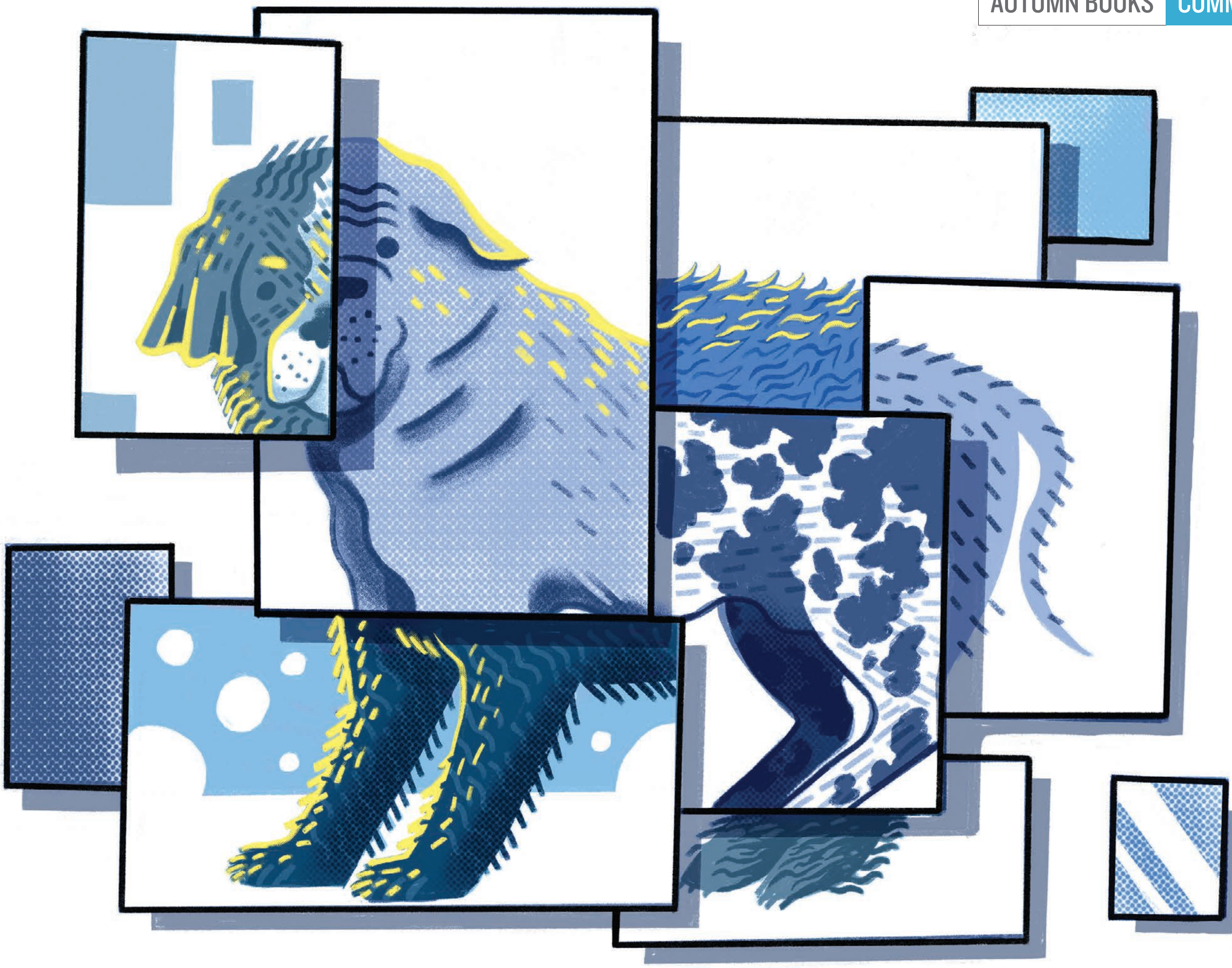
The authors' exhaustive documentation of these socio-economic forces supports their thesis that today's dog is a deliberate invention of the Victorians. But, for all the research and reporting, they do not explain the emotional

NEW IN PAPERBACK

Highlights of this season's releases.



A Million Years of Music
Gary Tomlinson MIT PRESS *(2018)*
"Musical expression is a universal characteristic of our species." Musicologist Gary Tomlinson explores the reaches of that idea, and to what extent the traits essential to music-making can be seen as evolutionary behaviours, traceable across human history. Expertly interweaving humanities and science, Tomlinson demonstrates how the answers to philosophical questions surrounding modern music can be discovered in their ancient origins.



drive behind it. Why did the Victorians want this dizzying variety? What did all that messing about with dogs' appearance do to the animals' emotions and behaviour?

Thanks to Russian geneticists Dimitry Belyaev and Lyudmila Trut (as chronicled in Trut and Lee Dugatkin's 2017 *How to Tame a Fox (and Build a Dog)*), we know that selecting for tame behaviour in wild foxes changes the animals' look. The reverse should also hold: that selecting for coat and eye colour manipulates genes that affect behaviour. If such bundled temperamental effects were mentioned in the vast nineteenth-century breeding literature, it would have been fascinating and important to include them.

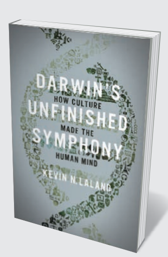
Given the profound sense of attachment on which the human–dog bond evolved, I would have expected that this bond — or the lack of it — would be an important factor in the Victorian explosion in breeding. We do learn that members of the Fancy were said to treat dogs better than their families; that toy dogs were bred to tolerate the excessive fawning of their mistresses; and that the shows' popularity was "driven by the participants' passion for dogs". The only evidence of passion, however, appears in those (mainly women) who fought to abolish the shows because of abusive conditions (such as long hours, lack of water and worse). They also spoke out against breeding standards that led to gross deformities and diseases still with us. Cruel ear cropping was abolished in Britain in 1895, but the 2007 ban for tail docking still allows exemptions for working breeds such as spaniels, poodles and pointers.

To me, the greatest service offered by *The Invention of the Modern Dog* is to remind us not to breed dogs for conformation alone. We knew that 150 years ago. Take the Dalmatian, which owes its spots to a gene profile associated with a painful urinary disease. A simple outbreeding to an English pointer in 1973 left the breed with spots and good health. In 2011, 15 generations later, the American Kennel Club finally recognized it as a true Dalmatian.

We now have the genetic science and technology to make true improvements to the twenty-first-century dog. We can and we must use this knowledge to re-invent the Victorian canine into an animal bred for good health and temperament. I can't wait to see what that dog looks like. ■

Meg Daley Olmert *is the author of* Made For Each Other: The Biology of the Human-Animal Bond *and director of research for Warrior Canine Connection, a service-dog training intervention to reduce the symptoms of combat trauma, in Boyds, Maryland.*
e-mail: meg@warriorcanineconnection.org

How to Tame a Fox (and Build a Dog)
Lee Alan Dugatkin and Lyudmila Trut
Univ. Chicago Press *(2018)*
Biologists Lee Alan Dugatkin and Lyudmila Trut chronicle Trut's extraordinary, long-running research with Dmitri Belyaev on the domestication of silver foxes — work that effectively shrank 15,000 years of evolution to decades.



Darwin's Unfinished Symphony
Kevin N. Laland Princeton Univ. Press *(2018)*
How did the human potential for culture evolve from hominin behaviour and cognition? Evolutionary biologist Kevin Laland navigates the false leads and breakthroughs that led to his theory that culture is both a result of evolution, and a factor that has effectively shaped its progress.

MEDICAL RESEARCH

Riding the global tide of blood

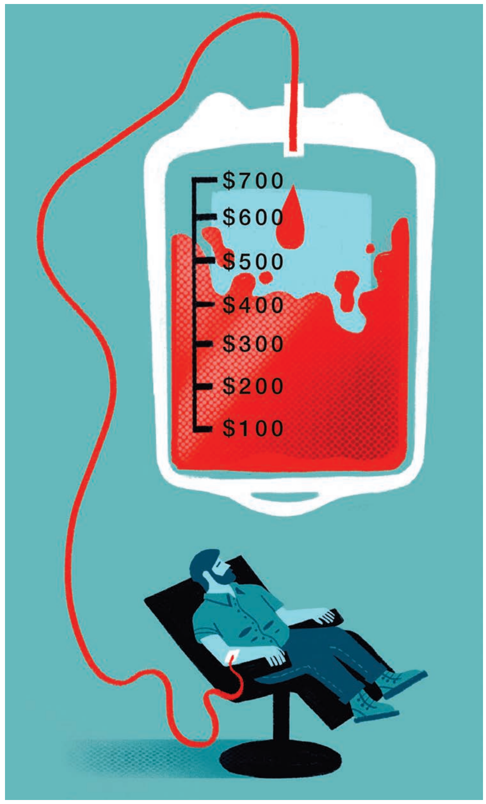
Tilli Tansey enjoys a compelling exploration of the dynamic biological fluid.

Blood is life. Blood is death. Writer Rose George's book ranges extensively and often disturbingly between these contradictory extremes. George examines blood as a life-saving medicine, an infective agent, an easily accessible indicator of disease and injury, a taboo, a weapon and, in all contexts, a commodity to be bought, sold, used, misused or controlled.

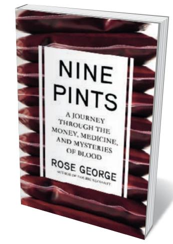
George (whose previous books examined shipping and human waste) develops each theme in a series of engaging personal stories and journeys. The "vein to vein" account of blood transfusion starts at St George's Hospital in South London, where George donates 470 millilitres of blood. She then 'follows' it to a National Health Service Blood and Transplant (NHSBT) processing facility in south-west England. There, a single donation can provide a range of products. These include red blood cells, platelets and cryoprecipitate for clotting disorders, as well as whole blood depleted of white blood cells to transfuse into infants with less-developed immune systems, and fresh frozen plasma for transfusion to replace lost blood. We learn that, since 2003, a "male donor preference" has operated in Britain: women's plasma, laden with excess hormones, mainly from contraceptive pills and hormone-replacement therapy, requires considerably more screening and treatment, and is routinely discarded.

George shares some sobering statistics. Every three seconds, someone in the world receives a blood transfusion (this translates to 2.5 million units of blood transfused per year in Britain, and 16 million in the United States). But many nations, including all those in Africa, fail to reach the World Health Organization's target of 1–3% of the population donating. In Sweden, a modest initiative begun in 2012 has increased contributions by simply texting donors to let them know when their blood has been used; Britain has followed suit.

But these life-saving donations can also carry death and disease. The disturbing



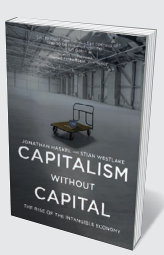
stories of contamination from around the world, especially by HIV and hepatitis C, are now well known. For instance, many people with haemophilia, surgical patients and new mothers who received blood products in the 1970s and early 1980s in Britain also unknowingly received infections, mainly HIV and hepatitis C. Many are still seeking recognition and compensation. The introduction of more-stringent criteria for donors has removed these problems from Britain's blood supply. There are effective measures. One is the rejection of blood donations from people who have recently visited areas where blood-borne diseases such as malaria, West Nile fever or Zika are rife. Another is testing for a wide range of viruses, including hepatitis B and C, HIV and syphilis. (Ironically, British blood is considered a risk for contamination with the prions causing the neurodegenerative condition Creutzfeldt–Jakob disease, and is not accepted outside the country.) Elsewhere, larger risks remain. HIV infection from a transfusion is 3,000 times more likely in India than in the United States. And, worldwide, as many as 10% of HIV infections have been calculated to come from blood products.



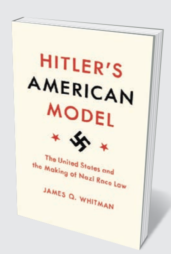
Nine Pints: A Journey through the Money, Medicine, and Mysteries of Blood
ROSE GEORGE
Metropolitan (2018)

There are further disturbing tales. We sense the terror of 16-year-old Radha in western Nepal: while she is menstruating, she must make her lonely way each evening to a remote hovel to sleep. This practice, *chaupadi*, makes Radha, like thousands of other teenagers, vulnerable to sexual assault by local men. George also recounts stories of poverty-stricken "plassers" in the United States, who legally sell their plasma twice a week (European limits are 24 donations a year) to earn US$2–3 per day. And there are the rural communities in India where blood is now a kind of cash crop. As George reports, this has led to horrific abuse of migrants, imprisoned as "blood slaves" and bled for cash.

Heroes and heroines, too, abound in *Nine Pints*. Janet Vaughan, considered "too stupid to be educated" by her headmistress, qualified in medicine in the mid-1920s and specialized in blood disorders at the London Hospital. She made crude liver extracts for the treatment of pernicious anaemia, using mincing machines borrowed from friends, including Virginia Woolf (a distant relative). By the late 1930s, war was looming, heralding a need for blood. Knowing of



Capitalism Without Capital
Jonathan Haskel and Stian Westlake
PRINCETON UNIV. PRESS *(2018)*
In our new industrial era, business assets are mostly intangible, from software to research. Jonathan Haskel and Stian Westlake offer insight into this mass economic shift, and how to exploit the move towards immaterial capital.



Hitler's American Model
James Q. Whitman PRINCETON UNIV. PRESS *(2018)*
Early US eugenics policies infamously helped to inspire Nazi atrocities. Here, legal historian James Whitman examines how US Jim Crow laws, which enforced racial segregation from the 1880s to the 1960s, became a model for the Reich's egregious anti-Semitic Nuremberg Laws.

advances made in collection and storage during the Spanish Civil War, Vaughan established several effective blood depots — one in a bar in Slough, which always attracted donors. She initiated a mobile service, using ice-cream vans to collect and deliver blood around the country. One contemporary commentator, Major General W. H. Ogilvie, considered the greatest medical advance of the Second World War to be not penicillin, but the blood-transfusion service.

For me, the outstanding hero is Arunachalam Muruganantham, an innovator in sanitary products from southern India. It is an area where menstruation is considered shameful and dirty, many women cannot afford commercially produced pads, and public toilets and running water are rare. The lack of basic hygiene and the use and reuse of inadequate washable rags can lead to girls and women missing out on education and employment, and contracting gynaecological infections.

Muruga, as he is known, noticed his wife using newspapers and cloth during menstruation, and decided to experiment with alternatives. He carried a football filled with goat blood under his clothes so that he could release the liquid as he moved, and gain some sense of the practical difficulties. Ridiculed even by his family, he persevered, and designed machines to manufacture affordable pads, encouraging local communes and factories to produce and sell them. Muruganantham's story has featured in a 2013 documentary by Amit Virmani, *Menstrual Man*, and a 2018 Bollywood feature film by R. Balki, *Pad Man* (see S. Priyadarshini *Nature* **555**, 27–28; 2018).

Nine Pints is highly readable and informative, but the chatty style grates at times, and there are a few irritating duplications. And the title — a nod to the volume of blood in a human body, which is variable and related to body size — seems strangely static for a dynamic biological fluid with many vibrant contexts. ■

Tilli Tansey *is emeritus professor of medical history and pharmacology at the William Harvey Research Institute of Queen Mary University of London.*
e-mail: t.tansey@qmul.ac.uk

PHYSICS

Black-hole fever

Richard Panek on two books tackling the counter-intuitive weirdness of these gravitational beasts.

Einstein's Monsters: The Life and Times of Black Holes
CHRIS IMPEY
W. W. Norton (2018)

Einstein's Shadow: A Black Hole, a Band of Astronomers, and the Quest to See the Unseeable
SETH FLETCHER
Ecco (2018)

In the late nineteenth century, physicist Ernst Mach wrote that when Isaac Newton published his theory of gravity in his book *Principia* (1687), it disturbed his fellow natural philosophers. The reason? It "was founded on an uncommon unintelligibility": two objects interacting without physical contact. Mach was trying to show how an affront to common sense gains respectability through familiarity. By his own era, gravity had become "common unintelligibility".

Black holes — gravitational beasts that warp space and devour light — have undergone a similar trajectory. In the 1980s, they still seemed like science fiction. Since then, advances in technology and theory have transformed them into scientific (near) certainties. Now, two books — *Einstein's Monsters* by astronomer Chris Impey, and science journalist Seth Fletcher's *Einstein's Shadow* — trace that transition without losing sight of how weird their subject is.

In *Einstein's Monsters*, Impey provides a history of black holes and an overview of investigations into their supremely counter-intuitive behaviour. The possibility of their existence arose from the idea that gravitation is a force of attraction between bodies of matter. If light were matter, as British philosopher John Michell and French mathematician Pierre-Simon Laplace argued in the eighteenth century, it would be subject to Newton's laws. And if Newton's laws were correct, an object with sufficient mass could overwhelm light's mass, creating a "dark star". Laplace even provided a mathematical foundation for such a thing, in 1799. That year, however, polymath Thomas Young demonstrated that light acts as a wave. Laplace dropped his idea.

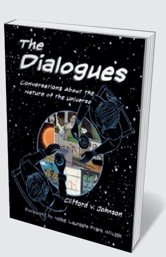
Albert Einstein's 1905 paper on the photoelectric effect, suggesting that light travels as both waves and packets of matter (photons), might have revived the dark star — had he not, ten years later, rendered obsolete the idea of gravitation as a force mysteriously operating without physical contact. In Einstein's universe, light follows curves in space-time created by the presence of objects with mass. Within months of Einstein's 1915 presentation of his general theory of relativity, astrophysicist Karl Schwarzschild found a solution for Einstein's equations: an object needn't be huge to trap light, as Michell and Laplace had assumed; it just needs to be sufficiently dense.

Impossibly so, thought many physicists, including Einstein. Such an object could result only from mass collapsing into a state of infinite density — a singularity. And infinities don't lend themselves to enthusiastic scientific endorsements. Just because a "monster" is mathematically feasible doesn't mean it exists.

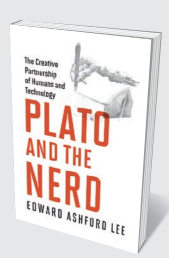
However absurd, the possibility lurked, and some theorists love lurking absurdities. From the 1920s, they had the benefit of quantum mechanics, an understanding of the subatomic Universe in which previously unimaginable density makes sense. Theorists in the 1930s calculated that the mass of a star determines its eventual fate — and that those fates include a neutron star or a dark star.

A neutron star was fine: quantum mechanics could account for a creature wherein a few cubic centimetres of matter weighs one billion tonnes, and the nuclei of adjoining atoms abut one another. A dark star was different. The ugly, infinity-dependent singularity at its heart defeated both general relativity and quantum mechanics.

For black holes to become commonly ▶



The Dialogues
Clifford V. Johnson MIT PRESS *(2018)*
Physicist Clifford Johnson aims to spark curiosity about science through discourse. In this thought-provoking graphic novel, he optimizes the rule of 'show, don't tell', encouraging engagement with concepts from the multiverse to immortality in everyday social scenarios.



Plato and The Nerd
Edward Ashford Lee MIT PRESS *(2018)*
Technology and creativity are irrefutably intertwined. Computer scientist Edward Ashford Lee explores both the potential impact of the digital revolution on human evolution and how technology's "real power comes from partnership with humans".

unintelligible, observations had to catch up. From mid-century, astronomers looked at the Universe in electromagnetic wavelengths from radio waves to γ-rays, and identified distant objects that generated geysers of radiation — a match for theoretical black holes. Theorists such as Stephen Hawking tried to divine what happens at the event horizon — the boundary between this Universe and whatever lies beyond the gravitational field.

Further advances, such as the Hubble Space Telescope, made the evidence overwhelming. Supermassive black holes probably occupy the centre of every galaxy and determine galactic growth. Over about 50 years, black-hole studies have gone from obscurity to a thriving industry. Theorists, Impey writes, are "in a golden age", and observers "are harvesting massive black holes on an industrial scale".

The harvest of two black holes is the subject of Fletcher's book. One lies in the relatively nearby Virgo A galaxy. The other, the supermassive candidate Sagittarius A*, is at our Galaxy's heart. Observations of black holes have generally relied on indirect evidence, given the constraints of attempting to 'see' a black object on a black background at distances of up to several billion parsecs. The evidence for Sagittarius A* includes numerous studies over the past 20 years, revealing the zigging and zagging of nearby stars and gas under its apparent influence. But the Event Horizon Telescope (EHT) has tried to observe it directly.

Fletcher, chief features editor at *Scientific American* (which shares a publisher with *Nature*), tells this story. To bring such an observation into the realm of the possible, a telescope would need an aperture the diameter of Earth. By using very-long-baseline interferometry — combining observations from multiple, far-flung radio telescopes — the EHT team conceived an apparatus effectively covering the Western Hemisphere. For one week in April 2017, that network focused on the centre of the Milky Way to extract images such as the blazing radiation that should be generated by matter heating up to billions of degrees as it orbits the black hole at velocities approaching the speed of light.

Fletcher secured close access to the EHT collaboration, particularly director Sheperd Doeleman. Its results aren't public yet, leaving a hole at the heart of Fletcher's narrative. He compensates with a compelling behind-the-scenes story of scientists struggling as much with funding and competition as with the challenges of seeing Sagittarius A*.

Both books address the seeming absurdities of their subject with authority and wit. Fletcher characterizes the EHT as a "distributed Babel, constructed on as many as a dozen high perches". And after describing a death spiral between two black holes, each 10 million times the mass of Earth and hurtling around each other at half the speed of light, less than 200 kilometres apart, Impey concludes: "This isn't an orbit, it's insanity."

Maybe. But if history is any guide, it won't seem so for long. Improvements to gravitational-wave detectors such as the Laser Interferometer Gravitational-Wave Observatory should make the detection of black-hole collisions routine, inspiring a new generation of theorists to address the incompatibility of general relativity and quantum mechanics. As these two books make clear, the study of black holes has progressed rapidly from "No way!" to "Oh, wow." The next step is: "What now?" ■

Richard Panek *is the author of* The 4% Universe *and the forthcoming* The Trouble With Gravity.
e-mail: richardpanek@yahoo.com

NEUROIMAGING

The brain decoders

Chris Baker enjoys a clear-eyed account of the promise and pitfalls of brain imaging.

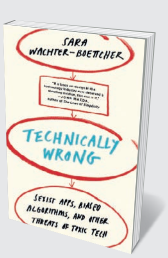
Since the advent of neuroimaging in the 1980s with positron emission tomography (PET), the sight of a living human brain in action has captivated scientists and the public. The emergence of functional magnetic resonance imaging (fMRI) in the early 1990s was a watershed. MRI scanners were already common in hospitals and, unlike PET, fMRI does not expose people to radioactivity. By measuring activity in the brain at the scale of a few millimetres, these scans seem to promise profound insight into the workings of the brain. That has led to wild claims that the technique could enable mind reading — actually knowing a person's precise thoughts.

The New Mind Readers: What Neuroimaging Can and Cannot Reveal about Our Thoughts
RUSSELL A. POLDRACK
Princeton University Press (2018)

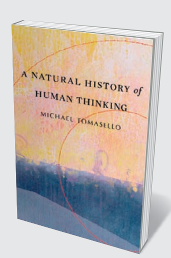
Russell Poldrack tackles these claims head on in *The New Mind Readers*. The experimental psychologist and neuroimaging pioneer takes readers through three decades of fMRI, its promise and limitations. From the race between groups in Minnesota, Massachusetts and Wisconsin in 1991 to show that MRI measures of blood oxygenation can reflect functional brain activity, to the development of techniques for decoding what someone is looking at, Poldrack surveys the history and biological basis of the technique and its potential application in areas as diverse as law and psychiatry.

Poldrack is an ideal guide. As director of the Stanford Center for Reproducible Neuroscience in California, he actively advances fMRI methods. His enthusiasm for them is clear, as is his frustration at how their data have been misinterpreted and abused.

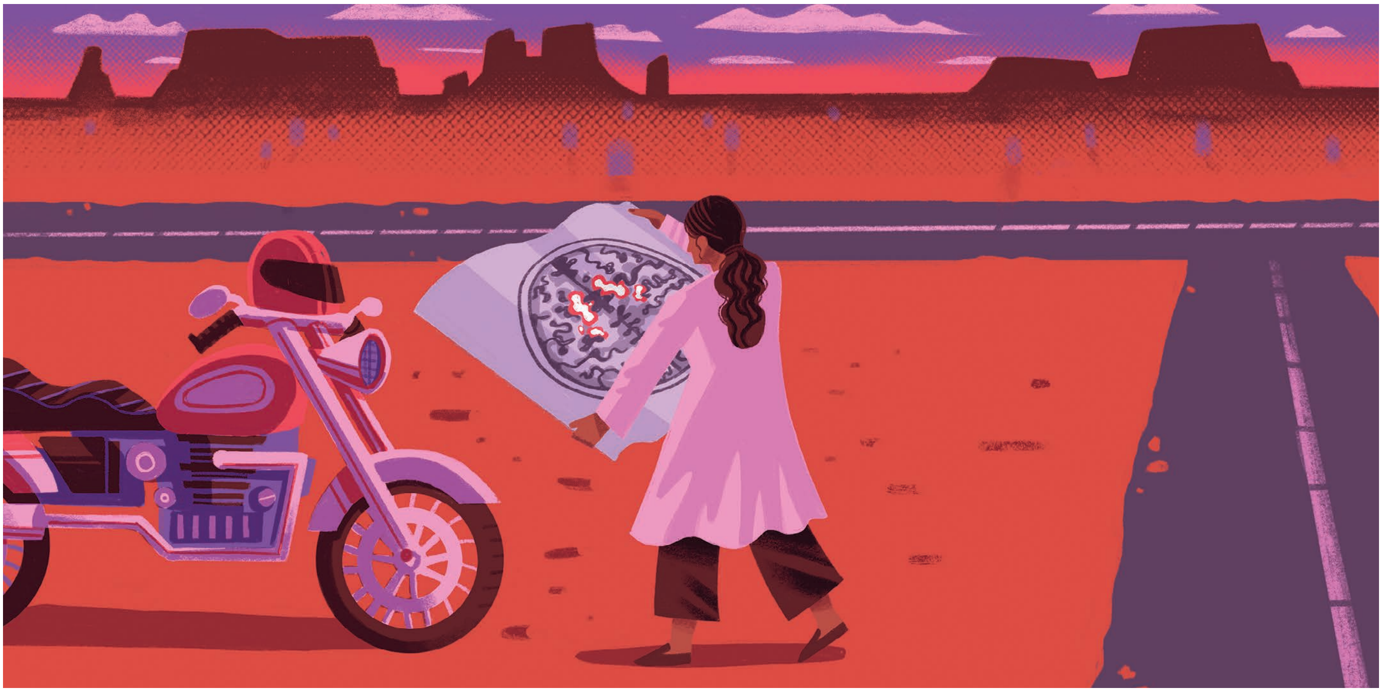
The technique has revolutionized neuroscience. Thousands of fMRI studies are published each year on topics ranging from perception to decision-making. For example, we now know that the pattern of blood flow



Technically Wrong
Sara Wachter-Boettcher W. W. Norton *(2018)*
Technology permeates life, from grocery shopping to dating apps. Yet we rarely question its design or aims. Web consultant Sara Wachter-Boettcher proffers a damning critique of the ethical dilemmas it poses, and why we need to demand more accountability from tech creators.



A Natural History of Human Thinking
Michael Tomasello Harvard Univ. Press *(2018)*
Drawing on 20 years of comparative studies on humans and great apes, psychologist Michael Tomasello theorizes that human cognition arose from social cooperation. Language and culture, he posits, also grew from our ancestors' need to work collaboratively.



to the fusiform face area in the temporal lobe can indicate that a person is looking at a face instead of a ball; and that imagining playing tennis or walking around your house, say, elicits activations in different brain regions. That is a major advance for neuroscientists and physicians who work with people in apparently non-responsive states after brain injury. It means they can identify patients with conscious awareness simply by asking them to engage their imaginations.

But some claims for fMRI are exaggerated. In 2007, *The New York Times* published an article based on fMRI data collected while people viewed images of candidates in US presidential primary elections, such as Barack Obama and John McCain. A group of neuroscientists at the University of California, Los Angeles, and political scientists had interpreted the results, alleging that they revealed how swing voters felt about the candidates.

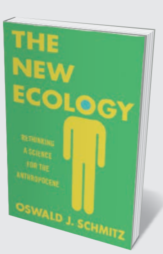
As Poldrack explains, the trouble is that activations of particular brain regions — such as the amygdala and the insula, which have been associated with fear and disgust, respectively — are not uniquely associated with particular mental states. One region, the anterior cingulate cortex, was found to be active in about one-quarter of thousands of studies that Poldrack and his colleagues examined, including those involving pain, short-term memory and cognitive control. So, 'reverse inference' of thoughts from brain-activation patterns can be very misleading. The potential for outlandish claims is high, Poldrack shows, when scientific data are used to support political and commercial interests — for example, when companies promote the ability to detect lies or to evaluate how viewers respond to advertising without sufficient scientific rigour. *The New Mind Readers* is a valuable example of how science can be discussed clearly and even-handedly, without sensationalism.

One of Poldrack's key themes is that interpreting fMRI findings demands an understanding of the underlying data and how they were produced. These scans do not measure neural activity directly. They rely on changes in the magnetic properties of haemoglobin (depending on levels of oxygen), to reveal local differences in blood flow. These reflect neural activity and are associated with different mental states, such as increases in the activity of motor cortex while tapping the fingers. When a technique involves hundreds of thousands of measurements across the brain, it is challenging to distinguish between a real change and a chance observation. Concerns over reproducibility are prominent. Moreover, many experiments use only a small sample, of fewer than 20 participants, often university students, in a laboratory setting.

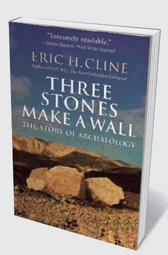
Caution is needed about generalizing to more complex, real-world situations. These include driving a car down a busy motorway, or moving from average activation patterns to single brains in a much more diverse general population, ignoring the importance of the individual variability that is part of being human.

The New Mind Readers is personal and selective. Poldrack gives short shrift to some methods, including brain stimulation, in which magnetic pulses are used to alter brain function directly to probe the ▶

THE POTENTIAL FOR OUTLANDISH CLAIMS IS HIGH.



The New Ecology
Oswald J. Schmitz Princeton Univ. Press *(2018)*
As we strive for sustainability amid unprecedented global transition, ecology is evolving to encompass the interdependence of human agency and nature. Ecologist Oswald Schmitz calls for careful stewardship and conservation of biodiversity to foster ecosystem resilience.



Three Stones Make A Wall
Eric H. Cline Princeton Univ. Press *(2018)*
Archaeologist Eric Cline walks us through the fascinating history of his discipline, traversing civilizations and the globe. From the discovery of Tutankhamun's tomb in Egypt to the future of excavation as technology advances, this is an engaging introduction to a gripping field.

specific role of the region targeted. He also skimps on neuroimaging techniques such as magnetoencephalography, which directly measures changes in magnetic fields produced by electrical signals in the brain. This is not an exhaustive account, and Poldrack focuses only on key developments and pioneers close to his own work. Yet his idiosyncratic approach is deeply engaging.

I was fascinated by Poldrack's description of why he decided to scan himself more than 100 times over 18 months to investigate how the brain changes over time — despite enduring a panic attack the first time he went into an MRI scanner. This intensive study uncovered much about the stability of brain function and the factors that affect it (including caffeine, food and mood). Yet Poldrack reveals that he learned "depressingly little" about himself during the experiment, highlighting the challenges of using fMRI for personalized medicine.

At times, Poldrack loses focus. His brief forays into topics such as the nature of mental illness are unsatisfying: they are too brief and lack the clarity of the rest of the book. Nevertheless, this is a compelling introduction that lucidly spells out the risks of taking media reports at face value, and urges readers to dig into the details. fMRI is evolving rapidly and researchers are just starting to map brain activity at sub-millimetre resolution, revealing activity — both in different regions and in different layers of cortex within a region.

Happily, despite the book's title, Poldrack makes it clear throughout that 'mind reading' as most people would imagine it remains in the realm of science fiction. What is much more exciting is the potential of fMRI for providing insight into brain function that will ultimately lead to clinical applications. ■

Chris Baker *is chief of the Section on Learning and Plasticity at the US National Institute of Mental Health in Bethesda, Maryland.*
e-mail: bakerchris@mail.nih.gov

The views expressed do not necessarily represent those of the US National Institutes of Health, the Department of Health and Human Services or the US Government.

ATOMIC PHYSICS

Secret histories of the bomb

Sarah Robey examines two books that together trace the birth and evolution of the nuclear age.

Fallout: Conspiracy, Cover-Up, and the Deceitful Case for the Atom Bomb
PETER WATSON
PublicAffairs (2018)

Burning the Sky: Operation Argus and the Untold Story of the Cold War Nuclear Tests in Outer Space
MARK WOLVERTON
Overlook (2018)

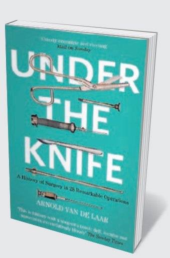
The secretive twentieth-century history of nuclear weapons is an evergreen subject. Writers have mined it for stories of breakneck innovation, wrenching controversy, unimaginable violence, espionage and larger-than-life personalities. Two new books — *Fallout* from historian Peter Watson and *Burning the Sky* by science writer Mark Wolverton — continue this trend, recalling two instructive episodes in our collective nuclear past.

Fallout synthesizes the history of the race to create an atomic bomb in Germany, the United Kingdom and the United States from the 1930s to the end of the Second World War — a story of duplicitous players, sinister decisions and regrettable outcomes. In a twist of historical fate, Adolf Hitler's rise coincided with major breakthroughs in particle physics, including the theorization of nuclear fission by Lise Meitner and Otto Frisch in December 1938. By the time war broke out, many prominent scientists had fled the Reich, and the Allies assumed that any physicists remaining in Germany, including Werner Heisenberg, were working to harness fission to produce a bomb (see A. Finkbeiner *Nature* **503**, 466–467; 2013). This was the main reason that Britain and the United States sought to beat Hitler to the punch.

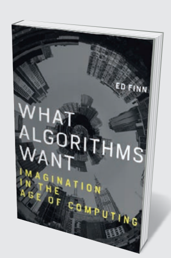
But, as Watson uncovers, British intelligence showed that Germany's atomic programme had stalled by 1942. Why, then, did the joint UK–US atomic programme move forward, despite incredible cost and danger? Watson painstakingly outlines a complex web of who knew what, and when, to show how a series of opportunities to stop what became the Manhattan Project arose, then passed. In 1942, without access to full British intelligence, the US government actually ramped up its project, assuming that Germany was advancing rapidly. As Watson puts it, "a series of momentous mistakes were made, and lies told" by French, German, British and US officials. Thus "the world stumbled, even blundered, unnecessarily into the nuclear age". In his view, today's extraordinary nuclear challenges — deteriorating arsenals, ongoing proliferation and the rebirth of sabre-rattling nuclear diplomacy — were preventable.

Watson's meticulous attention to this chronology is one of the book's strengths. He details wartime research on both sides of the Atlantic, from Copenhagen to New Mexico, and delves into the motivations and actions of the Allied leadership. Also interesting are his findings on the public availability of nuclear research in contemporary press reports and scientific journals, including *Nature*. Managing these threads is no small authorial feat — of research, especially. Watson also weaves together the insights of previous nuclear historians, such as Richard Rhodes, Martin Sherwin and David Holloway.

In what could have been a volume in its own right, the narrative is bookended by the overlapping wartime sagas of Niels Bohr and Klaus Fuchs. Fuchs, the infamous German



Under The Knife
Arnold van de Laar JOHN MURRAY *(2018)*
In this witty chronicle, surgeon Arnold van de Laar dissects thousands of years' worth of remarkably gruesome stories. From anaesthetic-free amputations and bloodletting to Albert Einstein's aneurysm, these are key insights into the cut and thrust of medicine.



What Algorithms Want
Ed Finn MIT PRESS *(2018)*
Algorithms saturate the digital universe, from Amazon book recommendations to Uber.
Ed Finn will make you reassess how you think about these formulae: not as mere components of code and computations, but shaped by a philosophy, and shaping culture in their turn.

spy who passed secrets from the Manhattan Project to the Soviet Union, and the towering physicist Bohr serve as foils to the main plot. During the war, each had a longer-term vision of what nuclear weapons would mean to the post-war order. Bohr suspected that the Western Allies' relationship with the Soviets would be damaged; Fuchs, by delivering crucial technical information to the Soviets, helped to ensure that it was.

The aftermath of that era comes to life in Wolverton's gripping *Burning the Sky*, the first book-length treatment of a remarkable series of nuclear tests in outer space, code-named Operation Argus. After the Soviet Union launched its satellite Sputnik-1 in 1957, US agencies realized that they were lagging behind in missile technology. The cold war arms race took on new urgency. When, in early 1958, the US government finally succeeded in launching its Explorer 1 satellite — the first of more than 90 in the series — the achievement did more than calm US anxieties. Experiments on board Explorers 1 and 3 led to the discovery of the Van Allen belts, concentrated bands of radiation that circle the planet along the contours of its magnetic fields.

At a time of international tensions and ample defence dollars, however, scientific discovery was rarely separate from weapons considerations. At Livermore Radiation Laboratory in California, physicist Nicholas Christofilos believed that the belts could be harnessed as part of US defence. He theorized that high-altitude nuclear detonations would create a "shell of radiation" that could destroy missiles and warheads. Convinced, the US government under President Dwight Eisenhower embarked on Operation Argus, and later Operation Fishbowl, to test the theory.

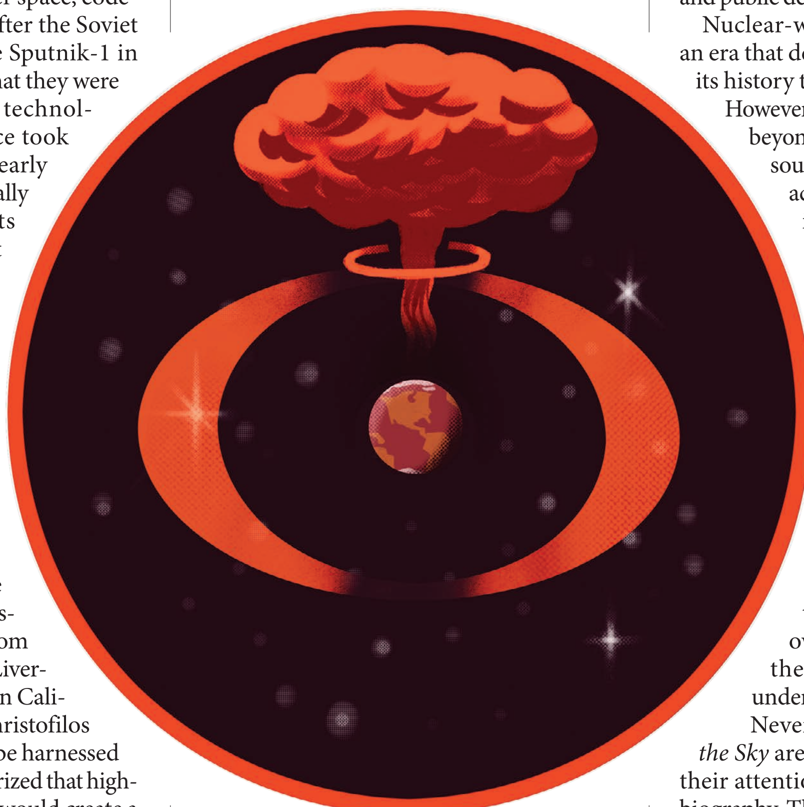
The tests were logistical nightmares. Argus was conducted in the remote South Atlantic amid treacherous weather and technical problems. Only the last of its three 1.5-kiloton weapons detonated at the projected altitude, 794 kilometres above Earth's surface, in September 1958. Half of the Fishbowl tests, in 1962, were aborted or cancelled. As Wolverton shows, it was incredible that there were no serious casualties. Although temporary belts were created, they were much weaker than Christofilos had theorized — capable of damaging satellites and releasing powerful electromagnetic pulses, but not of stopping a missile.

Together, these books highlight the tensions endemic to classified state-sponsored research in democratic society. Watson's subjects, including Manhattan Project heavyweights J. Robert Oppenheimer and Edward Teller, operated under — and were arguably victims of — extreme compartmentalization of information. Participants had a very limited view of the whole project. Fifteen years later, Wolverton's actors faced secrecy of a new kind. The Explorer missions took place under the auspices of the International Geophysical Year (IGY), the 1957–58 venture in which more than 60 countries shared data and took part in peaceful scientific collaboration. But as classified military studies, Argus ran counter to the IGY mission. Ultimately, national security lost out to scientific cosmopolitanism.

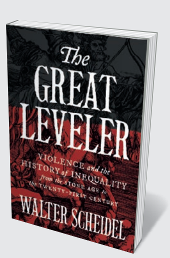
Whereas the Manhattan Project and its vast infrastructure of labs and production sites largely dodged press exposure, Argus officials struggled to contain leaks. By the late 1950s, the public was demanding nuclear transparency, and agencies involved in Argus, such as the Atomic Energy Commission, knew the tests would be controversial. The experiments were finally revealed in 1959 by *The New York Times*, inciting a media frenzy and public debate.

Nuclear-weapons science was born of an era that demanded secrecy. Recounting its history thus demands a sceptical lens. However, problems arise if it is assumed, beyond available evidence, that every source has a secret meaning, every actor an ulterior motive. Watson in particular has a penchant for the conspiratorial, and is eager to expose and blame. That makes for a page-turning read, but can discount the context of war — hot and cold. Both books are populated by egotists and opportunists, the quest for scientific priority, nationalism, ignorance, suspicion and doubts. Ultimately, their stories are all too human. Historians should not absolve their subjects entirely, but we owe it to past individuals, even the most belligerent, to try to understand all the forces at play.

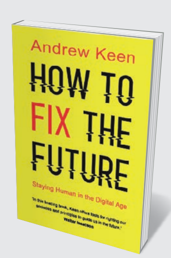
Nevertheless, *Fallout* and *Burning the Sky* are informative and balanced in their attention to diplomacy, science and biography. They also provide much to ponder concerning the state of play now, from the nuclearization of North Korea to the unknown future of the Iran nuclear deal, and the part that the first members of the exclusive nuclear club might have to play in future. ■



Sarah Robey *is assistant professor of the history of energy at Idaho State University in Idaho Falls. She studies and writes about nuclear weapons in US society.*
e-mail: robesar5@isu.edu



The Great Leveler
Walter Scheidel PRINCETON UNIV. PRESS *(2018)*
In this monumental, pessimistic study, historian Walter Scheidel examines anew an old social issue: economic inequality. As he reveals, disparities have burgeoned during times of peace, declining only during wars and revolutions. "Inequality never dies peacefully," he notes.



How To Fix The Future
Andrew Keen ATLANTIC *(2018)*
The Internet has advanced from a communication device to an unstoppable force moulding societies. Andrew Keen, pioneer of the cyber-tsunami, uses lessons from the Industrial Revolution to envision a future relationship with life online that honours human values. **Mary Craig**

Correspondence

Close loophole for chemical weapons

As the Fourth Review Conference of the Chemical Weapons Convention meets next month, state parties need to address mounting concerns about the potential development and use of law-enforcement weapons involving chemical agents that act on the central nervous system (CNS).

Since 2013, when the Organisation for the Prohibition of Chemical Weapons was awarded the Nobel Peace Prize for ridding much of the world of stockpiled chemical weapons, lethal nerve agents have been used in Syria (sarin), Malaysia (VX) and the United Kingdom (novichok).There is a high risk that our enhanced understanding of the brain, coupled with rapidly advancing technology, will facilitate the development of increasingly dreadful chemical weapons.

Article II.9(d) of the Chemical Weapons Convention designates law enforcement, including domestic riot control, as a potentially acceptable purpose for the use of certain toxic chemicals (provided that the types and quantity used are consistent with this purpose). However, the range of potentially permissible chemicals has not been established. This provides a possible loophole for states to use CNS-acting chemicals for law enforcement. The use and development of ever-more sophisticated agents for such purposes would work against the prohibition of chemical weapons.

We strongly believe that this potential loophole must be closed. There are 39 countries that publicly support an initiative led by Australia and Switzerland against the use of such aerosol agents in law enforcement. In our view, a crucial first step is for the meaning and application of the convention in this area to be clarified at the review conference so as to ensure that CNS-acting chemical agents cannot be used for law-enforcement purposes.

Lijun Shang, **Michael Crowley**, **Malcolm Dando** *University of Bradford, UK.*
m.r.dando@bradford.ac.uk

Co-producers: open data can test trust

I agree that trust and open data are essential for successful collaborations between stakeholders and scientists (see go.nature.com/2quegjd and, for example, *Nature* **562**, 7; 2018). However, what happens to raw data once they become freely available can erode participants' trust in science — as I found when working with farmers in a pilot survey of soil health earlier this year.

In this survey of more than 1,300 hectares in the United Kingdom, farmers monitored earthworm populations on their land. Earthworms are good indicators of farmland biodiversity. The data will help to underpin initiatives such as the ongoing national #30minworms farmland survey, which aims to make crop production more sustainable (see go.nature.com/2owgztz).

Raw data are openly available for other earthworm surveys conducted at the Broadbalk field-trial site in Harpenden, UK, which has records going back to 1843 (see go.nature.com/2yz8u8q). An independent analysis of these unreplicated data concluded that earthworm populations are in drastic decline (R. J. Blakemore *Soil Syst.* **2**, 33; 2018). Unfortunately, the paper prompted alarming media speculation, such as: "Farmers around the world have been turning verdant fields into subterranean deserts" (see go.nature.com/2ebruek). Many of the farmers I'd worked with were shocked to see those data, painstakingly collected like their own, trivialized in this way by the media.

Open data sets assembled from participatory science must not be seen as a liability by research co-producers. I suggest that publishers could help to protect trust between research co-producers by developing best-practice guidelines specifically for these data sets.

Jacqueline L. Stroud *Rothamsted Research, Harpenden, UK.*
jacqueline.stroud@rothamsted.ac.uk

Policy training for junior researchers

Early-career researchers can be promising candidates for informing and shaping science policy (*Nature* **560**, 671–673; 2018). Given the necessary support, they could learn to engage with policymakers and to create sustainable interactions with them for the future.

Senior researchers would need to share their knowledge and networks with these new team members. Research institutions could offer regular training — or even integrate it into the curriculum. Debates on policy implementation strategies, stakeholder involvement and far-reaching changes in science-policy systems might all be included.

Once active at the science–policy interface, early-career researchers would be in a position to inspire and mentor their peers to follow them.

Norma Bethke, **Paul Gellert**, **Joachim Seybold** *Charité–Universitätsmedizin Berlin, Berlin, Germany.*
norma.bethke@charite.de

Junior researchers need a break, too

What does your out-of-office reply say when you are on holiday? Chances are, your answer depends on whether you are a junior researcher or a tenured professor.

During the summer break, I sent an e-mail through the mailing list of the German Society for Psychology, and received 223 out-of-office messages in response. Most replies (150) did not specify whether e-mails would be read or not, or if the sender was out of office for maternity leave or illness. For those who indicated that e-mails would not be read at all (31 replies), it was almost twice as likely that the message came from a professor (21 replies, or 68%) than when the message indicated that e-mails would be read occasionally (42 replies, of which 34% came from professors). Junior researchers' replies (37 in total) fell predominantly into the latter category, with just 10 saying e-mails would not be read.

Given the technological possibilities for accessing our e-mail accounts even in the remotest corners of the world whenever we feel like it, abstinence is a deliberate choice. This choice is apparently easier for a tenured professor to make than it is for a junior researcher.

I suggest that, as a community, we should create an environment in which the choice over whether to read e-mails during holiday periods is not dependent on seniority (see also J. Overbaugh *Nature* **477**, 27–28; 2011).

Jan Philipp Röer *Witten/Herdecke University, Witten, Germany.*
jan.roeer@uni-wh.de

Nature readers can cope with faeces

I was surprised by your Research Highlight 'Why naked mole rats eat poo' (*Nature* **561**, 9; 2018). It was not so much the content that surprised me but the use of the colloquial and decidedly juvenile headline. I hesitate to speak on behalf of the global scientific community, but I think it's safe to assume that we have the nous and maturity to deal with the word 'faeces'.

Stephen E. Moss *University College London Institute of Ophthalmology, London, UK.*
s.moss@ucl.ac.uk

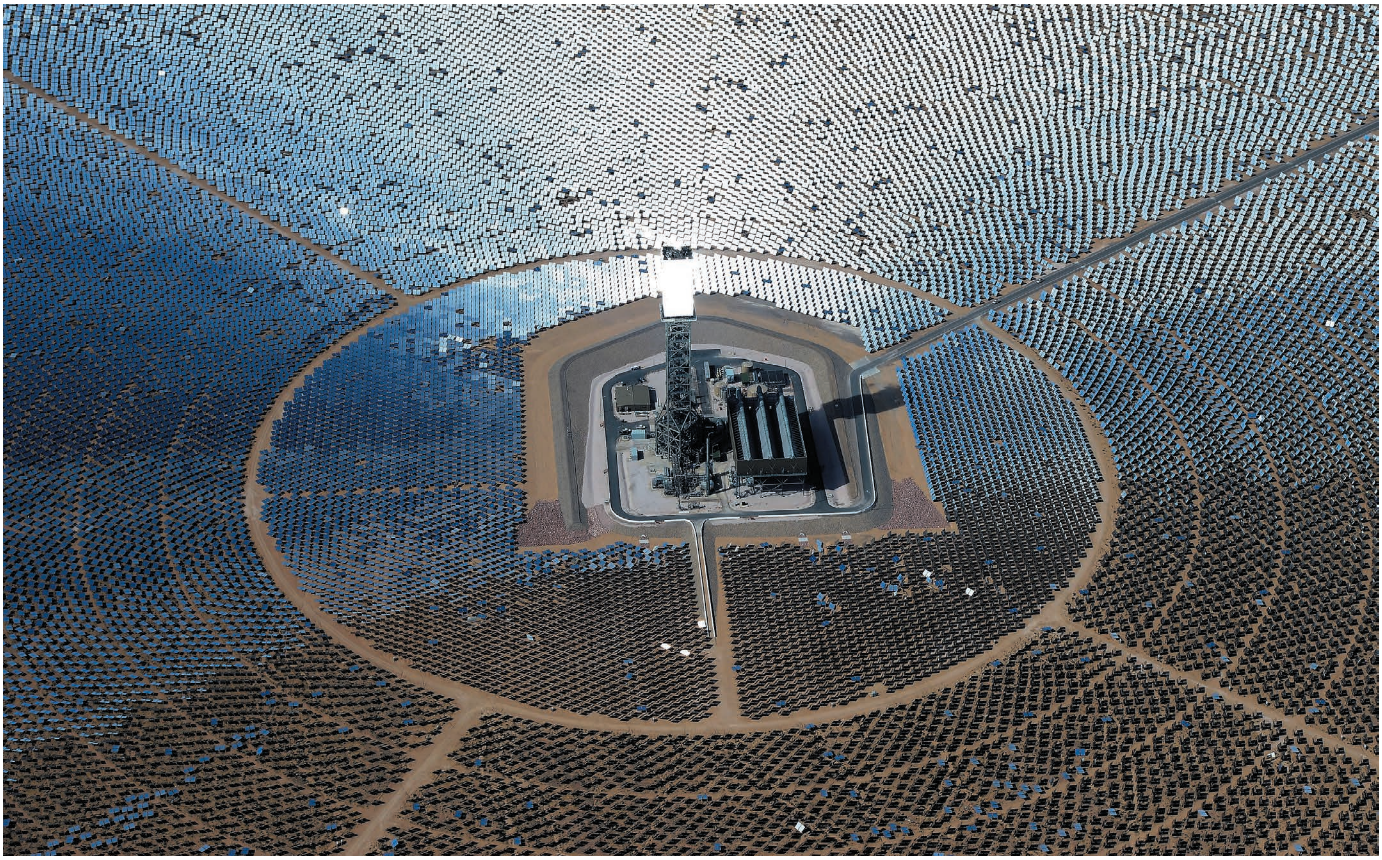

ETHAN MILLER/GETTY

Figure 1 | The Ivanpah Solar Electric Generating System, Las Vegas. Concentrated solar power plants such as this one generate heat by focusing sunlight onto a central tower using mirrors. The heat is then used to drive power cycles for electricity production. Caccia *et al.*[1] report a metal–ceramic composite material designed to make high-temperature heat exchangers — devices needed to enable power cycles in future concentrated solar power plants.

MATERIALS SCIENCE

A composite that takes the heat

A remarkable metal–ceramic composite material has been produced that could aid the development of the next generation of power plants — and might even have a role in curing the world of its addiction to fossil fuels. SEE LETTER P.406

CRAIG TURCHI

Composite materials that combine metals and ceramics have been developed for many different applications, such as use in wear-resistant surfaces for tools and engine parts, electrical components and even dental fillings. On page 406, Caccia *et al.*[1] report a metal–ceramic composite with a combination of properties that makes it suitable for a very different application — in devices known as heat exchangers, which must work at high temperatures in power plants. By enabling highly efficient heat transfer, the new material might allow the realization of a cost-effective electricity-generation process that is currently being developed based on a fluid phase of carbon dioxide known as supercritical CO_2.

Metals and ceramics have been around for centuries, and have their own distinctive properties and applications. For example, bronze and iron have good shock resistance and are malleable enough to be worked into complex shapes such as helmets and horseshoes. Ceramics, such as the materials used to make pottery, can be formed into simple shapes and are prized for their resistance to heat and corrosion. These two classes of material have therefore found disparate applications, and for a long time marched along separate technological paths.

In the mid-twentieth century, the advent of jet engines generated a need for materials that had high resistance to heat and oxidation, the ability to cope with rapid temperature changes, and excellent mechanical strength, exceeding the properties of available metals. The US Air Force funded research to make ceramic–metal composites that had these properties, and the word 'cermet' was coined to describe them. Cermets have since been developed for multiple applications, but, in most cases, they have been used for small parts or surfaces. Caccia *et al.* now report a cermet that can withstand extreme temperatures, high pressures and rapid thermal cycling.

To make this cermet, the authors first produced a preform — a precursor that needs further processing to be turned into the required final object, analogous to the unfired version of a clay pot. The authors compacted tungsten carbide (WC) powder into the approximate shape of the target object and heated it at 1,400 °C for 2 minutes to bond

the particles together. They then machined the porous preform to generate the desired final shape.

Next, the authors heated the preform in a chemically reducing atmosphere (a mixture of 4% hydrogen in argon) at 1,100 °C and immersed it in a vat of liquid zirconium and copper (Zr_2Cu) at the same temperature, before removing it and finally heating it at 1,350 °C. This process causes the zirconium to displace the tungsten from the tungsten carbide, producing zirconium carbide (ZrC), tungsten and copper. The liquid copper is forced out of the ZrC matrix as the material solidifies, so that the final object is formed of approximately 58% ZrC ceramic and 36% tungsten metal, with small amounts of residual tungsten carbide and copper. The beauty of the method is that the porous preform is converted into a non-porous ZrC/tungsten composite of the same dimensions (the overall volume change is approximately 1–2%).

The clever manufacturing process is complemented by the robust properties of the final product. Caccia *et al.* find that, at 800 °C, the ZrC/tungsten cermet conducts heat 2.5–3 times better than iron- or nickel-based alloys currently used in high-temperature heat exchangers — which should improve the effectiveness of such devices. Furthermore, the mechanical strength of the ZrC/tungsten cermet is higher than that of nickel-based alloys typically used in high-temperature applications, and is unaffected by temperatures up to at least 800 °C, even when the cermet had previously undergone 10 heating–cooling cycles between room temperature and 800 °C. By contrast, iron alloys (stainless steels) and nickel alloys lose 80% or more of their strength at temperatures between 500 °C and 800 °C (ref. 2).

Heat exchangers transfer the thermal energy generated by a power plant to the working fluid in a thermal engine (such as a steam turbine) that converts heat into mechanical energy. The mechanical energy in turn is used to generate electricity. The overall process of converting heat to electricity is known as a power cycle. The US Department of Energy, along with industrial partners, is currently building a 10-megawatt test facility for a power cycle that uses supercritical CO_2 as the working fluid (see go.nature.com/2pi50mt). This power cycle promises lower costs and greater efficiency for future power plants, compared with currently used power cycles, but requires highly efficient heat exchangers. Caccia and colleagues' paper focuses on heat exchangers that could be used in this power cycle in concentrated solar power plants (which use sunlight concentrated by mirrors to generate electricity; Fig. 1), but the heat exchangers could also be used in advanced nuclear and fossil-fuel-fired power plants.

One technical challenge that must still be addressed concerns the oxidation resistance of the new cermet: the material is prone to oxidation in air at high temperatures such as might be experienced in a power-plant heat exchanger. Supercritical CO_2 is only a weak oxidizing agent, but could still break down the cermet. Caccia *et al.* report that cermet oxidation can be prevented for up to 1,000 hours at 750 °C and at high pressure (20 megapascals) when the material is coated with a thin layer of copper, and if a small amount of carbon monoxide (50 parts per million) is mixed with the supercritical CO_2. Nevertheless, long-term durability must still be proved.

Lastly, the authors' preliminary estimates indicate that the combined costs of raw materials and processing required to make a heat exchanger from the ZrC/tungsten cermet would be lower than for an analogous heat exchanger made from a conventional nickel alloy. Moreover, the cermet device would provide twice the power density — that is, it could be half the size of its nickel-alloy counterpart. The use of such heat exchangers might help to reduce the costs of renewable concentrated solar power, making it economically competitive with fossil-fuel-derived electricity. ■

Craig Turchi *is in the Thermal Sciences Group at the National Renewable Energy Laboratory, Golden, Colorado 80401, USA.*
e-mail: craig.turchi@nrel.gov

1. Caccia, M. *et al. Nature* **562**, 406–409 (2018).
2. ASME Boiler & Pressure Vessel Code, Section II, Part D (2013).

BIOTECHNOLOGY

CRISPR tool puts RNA on the record

The bacterial-defence system CRISPR–Cas can store DNA snippets that correspond to encountered viral RNA sequences. One such system has now been harnessed to record gene expression over time in bacteria. **See Article p.380**

CHASE L. BEISEL

Determining the gene-expression profile of a cell is crucial to unlocking how its DNA blueprint gives rise to its physical characteristics and behaviours. The standard approach used currently involves RNA sequencing or single-cell imaging techniques that generate detailed snapshots of gene-expression profiles. However, these techniques capture such profiles only at the moment of analysis, and kill the cells. This makes it hard to capture fleeting gene-expression profiles or those that provide a complete picture of cells going through major behavioural or environmental changes. On page 380, Schmidt *et al.*[1] report progress in overcoming this challenge by enlisting a bacterial-defence system that can create a DNA record of specific RNA sequences in a cell.

The CRISPR–Cas bacterial-defence systems are probably best known for their application in genetic engineering to cleave specific DNA sequences[2]. But another feature of these systems is the incorporation of snippets of DNA from unwanted intruders into a bacterium's own genome. These stored sequences provide a permanent 'memory' of infection, which can enable a defensive response if the same sequences are encountered again. The nucleotides are added to the cell's DNA in a configuration called a CRISPR array. The sequence of an array alternates between identical repeat sequences and the incorporated snippets, which are called spacers. As spacers are acquired, the array lengthens, and the positioning of spacers in the array reflects the order in which they were inserted[3].

Almost all CRISPR–Cas systems acquire foreign genetic material by directly capturing DNA from an invader. Some previous work exploited this feature of CRISPR–Cas systems to record information in the form of acquired and stored nucleotide sequences. For instance, one approach[4,5] used CRISPR–Cas-mediated acquisition of externally provided synthetic DNA to capture sequences in a specific order. The particular order of the nucleotides in the spacers was subsequently 'decoded' to link each CRISPR array to pixels in sequential images[5]. Another study[6] used chemical cues from the environment to drive expression of a gene controlling the abundance of a form of circular DNA called a plasmid. As plasmid abundance rose in the cell, the plasmid became the preferred source of DNA snippets for new spacers; this linked the presence of the chemical cue to a stored spacer that matched the plasmid DNA. That study, in particular, set the stage for the use of CRISPR–Cas to record the expression of one or a few genes. Yet it was unclear how this approach could be extended to provide a comprehensive record of the gene-expression profile of a cell.

Schmidt and colleagues devised a creative solution by focusing on CRISPR–Cas systems that capture invading RNA rather than DNA[7] (Fig. 1a). These systems need only two proteins to achieve this feat, with one protein making a DNA version of the RNA sequence

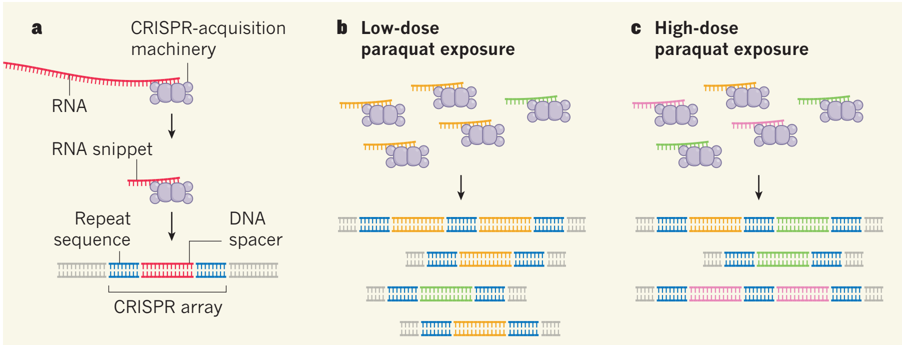


Figure 1 | A system to track RNA expression in cells. Schmidt *et al.*[1] report the development of a technique to monitor gene expression by storing, and subsequently sequencing, DNA sequences that correspond to the RNA sequences expressed in bacterial cells. **a**, The authors engineered *Escherichia coli* bacteria to express CRISPR-acquisition machinery proteins from the bacterium *Fusicatenibacter saccharivorans*. These proteins can capture RNA transcripts and cleave a short snippet of RNA that is used to make a DNA version of the sequence — a spacer. This spacer is incorporated between repeat sequences of DNA to form a CRISPR array. **b**,**c**, The authors tested whether such captured DNA profiles could be used to document the abundance of RNA transcripts generated in response to different conditions, such as different dosages of the toxic molecule paraquat. They found that the technique could capture a 'fingerprint' of the gene-expression profile that was characteristic of the specific condition encountered by the cell.

that becomes the spacer. Being able to generate DNA from RNA raised the possibility that this DNA could be used to document the identity and abundance of RNA transcripts, and therefore capture a cell's gene-expression profile.

To use these CRISPR–Cas systems, the authors first had to overcome two technological hurdles. One hurdle was finding efficient RNA-capturing Cas proteins, because previously characterized proteins were inefficient at this task. The authors tested a large and genetically diverse set of Cas proteins, and identified clear winners from the human gut bacterium *Fusicatenibacter saccharivorans*. The other hurdle was being able to conduct DNA sequencing that focuses on the few CRISPR arrays that had obtained a new spacer, because most arrays were unaltered. The authors overcame this hurdle by developing a simple approach that selectively isolates the CRISPR arrays that have newly acquired spacers.

With these advances made, the authors went on to develop a method they called Record-seq for capturing gene-expression profiles. They genetically engineered the bacterium *Escherichia coli* to contain the RNA-acquisition proteins from *F. saccharivorans*. They then verified that these proteins could incorporate spacers into the genetic information of the *E. coli* cell, and that RNA rather than DNA sequences determined the corresponding spacer DNA.

In Record-seq, the RNA-acquisition proteins are expressed during the recording of the gene-expression profile. At the end of this period, a sample of the cell population is taken. Newly expanded CRISPR arrays are isolated and sequenced, and the spacers are matched to the corresponding genomic sequences.

The next steps were to prove that the method could faithfully create a record of gene expression and to determine what could be discerned about the cellular environment during the recording period. The authors found that Record-seq could record hundreds to thousands of different RNA transcripts present in the cell at any time. Although there was a strong bias towards capture of highly abundant transcripts, the transcript abundance of particular RNAs, as assessed by RNA sequencing, generally correlated with the frequency with which a corresponding spacer sequence was acquired in the sample. Furthermore, the collection of spacers could form a particular pattern depending on the growth conditions in which the cells were cultivated, allowing the authors to use such a spacer 'fingerprint' as a way to discern the conditions that the cells had experienced.

One key outcome was that the authors determined the characteristics that seem to govern the selection of RNA snippets (typically averaging around 40 base pairs in length) by the CRISPR-acquisition machinery during the process that generates spacers. Schmidt and colleagues found that the snippets were rich in adenine and thymine nucleotides, and often came from either one of the two ends of an RNA transcript. Unexpectedly, the authors found no obvious preference for specific sequences flanking the RNA regions used to make RNA snippets. Such flanking sequences, often termed protospacer-adjacent motifs (PAMs), are needed for the recognition process that enables CRISPR–Cas defences to specifically cleave the intended target sequence in the invader but not to cleave the same sequence present in the array[8]. The system therefore might generate some spacers that will not enable an effective immune response to be launched because the corresponding target sequences are not flanked by a PAM. This possibility, and the ability of the RNA-acquisition proteins to acquire RNA snippets from the bacterium's own transcripts, raises questions about whether, and, if so, how, these systems effectively defend cells from unwanted intruders.

Arguably the most important demonstration of their method came when the authors compared Record-seq with direct sequencing of RNA. In one key experiment, the authors evaluated how well each technique could capture bacterial cells' transcriptional responses to a brief exposure to the toxic molecule paraquat. They found that only Record-seq could capture both transient and dosage-dependent features of the transcriptional response to paraquat exposure (Fig. 1b,c).

Schmidt and colleagues have laid the groundwork for using Record-seq to monitor complex gene-expression profiles over time, although there are some immediate technical limitations that must be overcome. One current limitation is that spacer acquisition still remains highly inefficient, requiring at least 10 million bacterial cells to faithfully record an expression profile. Another is that the authors tested their system only in bacterial cells, whereas much of the future potential of Record-seq might lie with animal and plant cells. Last, Record-seq was used to sequence arrays that have only one or two spacers, for reasons relating to how the newly expanded arrays were isolated and sequenced. If the technique is modified to analyse longer arrays, this could provide a way of discerning the timing and intensity of more than one cellular event during the same recording period. The successful application of DNA-based CRISPR technologies in various multicellular organisms, along with ongoing advances in the engineering of Cas proteins[9–11], offer hope that Record-seq might overcome these challenges and eventually provide a robust and widely used technology.

As Record-seq is further developed, it might have many applications. Could it be used to track spatio-temporal changes in gene-expression profiles in multicellular systems and shed light on the development of animal and plant tissues and organs? Perhaps microbial communities in fluctuating micro-environments or the interactions between a pathogen and its host during infection could be monitored using this technique. Finally, will it be possible to use cells engineered to perform Record-seq to monitor gene expression in difficult-to-access environments, such as the human gut, or to identify gene-expression profiles that are a signature of disease or abnormality? Schmidt and colleagues' technique might transform how gene-expression profiles are monitored *in vivo* in cells, and it highlights yet another aspect of CRISPR–Cas systems that can be harnessed to make powerful technologies. ■

Chase L. Beisel *is at the Helmholtz Institute for RNA-based Infection Research and the University of Würzburg, 97080 Würzburg, Germany.*
e-mail: chase.beisel@helmholtz-hiri.de

1. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. *Nature* **562**, 380–385 (2018).
2. Barrangou, R. & Doudna, J. A. *Nature Biotechnol.* **34**, 933–941 (2016).

3. Levy, A. *et al. Nature* **520**, 505–510 (2015).
4. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. *Science* **353**, aaf1175 (2016).
5. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. *Nature* **547**, 345–349 (2017).
6. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang H. H. *Science* **358**, 1457–1461 (2017).
7. Silas, S. *et al. Science* **351**, aad4234 (2016).
8. Leenay, R. T. & Beisel, C. L. *J. Mol. Biol.* **429**, 177–191 (2017).
9. Kleinstiver, B. P. *et al. Nature* **523**, 481–485 (2015).
10. Slaymaker, I. M. *et al. Science* **351**, 84–88 (2016).
11. Hu, J. H. *et al. Nature* **556**, 57–63 (2018).

This article was published online on 3 October 2018.

In Retrospect

Method for studying dark matter turns 25

In 1993, two papers reported observations of an astronomical phenomenon called gravitational microlensing. The results showed that microlensing could be used to probe the elusive dark matter that is thought to pervade the Universe.

GRZEGORZ PIETRZYŃSKI

One of the biggest mysteries in astronomy is the nature of dark matter, which is thought to account for about 85% of the matter and 25% of the total energy in the Universe[1]. There are several strong pieces of evidence for dark matter[2]. In particular, spiral galaxies such as the Milky Way have flat rotation curves[3] — graphs that show the orbital speed of stars as a function of their distance from the galactic centre. This property indicates that spiral galaxies are surrounded by large quantities of unseen matter. Dark matter has not yet been detected directly, so its identity is still unknown. But 25 years ago, Alcock *et al.*[4] and Aubourg *et al.*[5] reported observations in *Nature* that paved the way to a better understanding of its properties.

In 1986, the Polish astronomer Bohdan Paczyński suggested an observational test[6] to determine whether dark matter present in the halo of our Galaxy is composed of astronomical bodies such as small stars, brown dwarfs, neutron stars or black holes. Such bodies are intrinsically faint and would therefore be difficult to see in the Galactic halo from Earth.

According to Einstein's general theory of relativity, these massive compact halo objects (MACHOs) could act as lenses, focusing light and amplifying the observed brightness of stars in nearby galaxies (Fig. 1). This phenomenon, called gravitational microlensing, is sensitive to even low-mass lenses. Paczyński therefore proposed that, by monitoring stars in nearby galaxies, astronomers could look for microlensing events and verify whether MACHOs can account for dark matter.

Conceptually, this test is simple. But in practice, it requires millions of stars to be monitored over a period of years. The reason is that the microlensing events have an extremely low probability of being detected, because the star, the MACHO and the observer all need to be aligned. At that time, it was a challenge to observe such a huge number of stars, accurately measure their brightness and analyse the resulting data.

In October 1993, Alcock *et al.* and Aubourg *et al.* independently announced the first candidates for microlensing events caused by dark objects in the Galactic halo. Alcock and colleagues used a dedicated 1.27-metre-diameter telescope. The telescope was equipped with two charge-coupled device (CCD) cameras that had a field of view of 0.5 square degrees, which was considered large at that time. The authors spent a year monitoring 1.8 million stars in a nearby galaxy known as the Large Magellanic Cloud, and discovered one microlensing candidate. By contrast, Aubourg and colleagues used photographic plates that had a field of view of about 25 square degrees. They monitored 3 million stars in the Large Magellanic Cloud for more than three years, and detected two microlensing candidates.

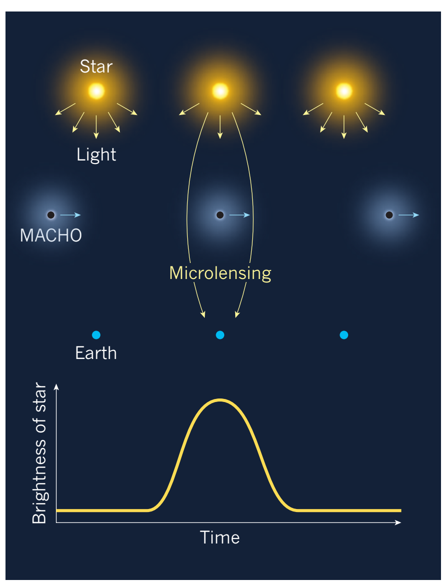


Figure 1 | Gravitational microlensing. In 1993, Alcock *et al.*[4] and Aubourg *et al.*[5] presented possible evidence for astronomical bodies called massive compact halo objects (MACHOs). Such bodies could account for dark matter — the 'missing' matter in the Universe. A MACHO can be detected when it passes in front of a star in a nearby galaxy. The MACHO bends light from the star towards Earth, which temporarily amplifies the star's observed brightness. This effect is known as gravitational microlensing.

The candidates all had fairly symmetrical light curves — plots that show the observed brightness of a star as a function of time (Fig. 1). Both teams obtained light curves in two different colours (blue and red) and found that the shapes of these plots were extremely similar. Such symmetrical and achromatic light curves are in agreement with what is expected for microlensing events.

However, the light curves of the two candidates obtained using photographic plates had a low signal-to-noise ratio. And for all three candidates, there was incomplete coverage of the brightening event, especially close to the peak brightness. Furthermore, although the frequency of the observed candidates was consistent with theoretical predictions, their low number prevented any firm confirmation that the light curves were products of microlensing — rather than light curves of a previously unseen class of astronomical object of varying brightness.

Despite these limitations, the potential discovery of microlensing by dark objects in the Galactic halo was spectacular. It demonstrated the feasibility of using microlensing to detect extremely faint stellar-mass or sub-stellar-mass bodies in the Galactic halo and, as a result, the feasibility of verifying whether dark matter in spiral galaxies is composed of such bodies. The microlensing surveys showed that it was possible to monitor millions of stars over years and to analyse the resulting enormous data sets, which was a breakthrough in observational astronomy.

Motivated by the work of Alcock *et al.* and Aubourg *et al.*, microlensing surveys improved enormously over the following 25 years, reaching capabilities to observe about 1 billion stars per night[7]. The huge data sets of high-quality light-curve data revolutionized many fields of astronomy — for example, the study of pulsating stars, extrasolar planets and star formation[8,9]. Moreover, millions of objects of varying brightness and thousands of microlensing events were detected, mostly in the direction of the Galactic bulge (a spherical structure near the centre of our Galaxy).

However, on the basis of observations of 35 million stars over eight years, only four microlensing events in the direction of the Large Magellanic Cloud were detected[10]. If these signals were caused by MACHOs, the contribution of these objects to the mass of the Galactic halo must be low (a few per cent). But the more likely explanation of these detections does not involve MACHOs at all, and relies on the lensing of stars in the Large Magellanic

Cloud by other objects in this galaxy.

Either way, MACHOs cannot account for all of the dark matter in spiral galaxies, and the identity of this mysterious matter remains unknown. The microlensing experiments ultimately gave a negative result. However, they have had a huge impact on many fields of modern astrophysics and have provided a lot of excitement and stimulation for the whole astronomical community. ■

Grzegorz Pietrzyński *is at the Nicolaus Copernicus Astronomical Center, 00-716 Warsaw, Poland.*
e-mail: pietrzyn.at.camk.edu.pl

1. Planck Collaboration. *Astron. Astrophys.* **594**, A13 (2016).
2. Trimble, V. *Annu. Rev. Astron. Astrophys.* **25**, 425–472 (1987).
3. Rubin, V. C., Ford, W. K. Jr & Thonnard, N. *Astrophys. J.* **238**, 471–487 (1980).
4. Alcock, C. *et al. Nature* **365**, 621–623 (1993).
5. Aubourg, E. *et al. Nature* **365**, 623–625 (1993).
6. Paczyński, B. *Astrophys. J.* **304**, 1–5 (1986).
7. Udalski, A. *EPJ Web Conf.* **152**, 01002 (2017).
8. Soszyński, I. *EPJ Web Conf.* **152**, 01001 (2017).
9. Gaudi, B. S. *Annu. Rev. Astron. Astrophys.* **50**, 411–453 (2012).
10. Wyrzykowski, Ł. *et al. Mon. Not. R. Astron. Soc.* **413**, 493–508 (2011).

NEUROSCIENCE

A living display system

Pigmented cells in the skin of cuttlefish can contract or relax to produce different skin-colour patterns. Tracking the dynamics of these cells reveals how this display system develops, and how it is controlled. SEE ARTICLE P.361

ADRIEN JOUARY & CHRISTIAN K. MACHENS

Our thoughts are hidden from sight, buried deep in the brain. Although this is undoubtedly beneficial in daily life, it is a serious drawback for neuroscientists: because much brain activity does not translate directly into behaviour, its function is difficult to determine. On page 361, Reiter *et al.*[1] take a step towards circumventing this problem. The authors studied cuttlefish, which can change their appearance on the basis of their perception of the external world — in essence, they display some of their 'thoughts' on their skin. Using a range of state-of-the art techniques for computer vision, spectrometry and biomathematics, together with electrophysiology, the group exposes one of the most complex systems of motor coordination ever recorded.

Cuttlefish, like squid and octopuses, are cephalopods. They have one of the largest brains among invertebrates, and can memorize complex spatial relationships or episodic events — abilities conventionally associated with mammals and birds[2]. These brainy molluscs lack a protective shell, but have evolved a sophisticated display system that enables them to quickly transform the colour and patterning of their skin in response to a changing perception of the world around them, generating a broad range of patterns used for camouflage, deception of prey or sexual communication[3].

The cuttlefish skin contains millions of cells called chromatophores, which can produce tiny dots of colour (yellow, orange, red, brown or black). If the radial muscles that control a chromatophore are relaxed, the pigments are imperceptible. But muscle contraction produces a colourful pixel several tens of micrometres wide[4] (Fig. 1). When viewed at a distance, the millions of individual pixels form a complex image in the style of a pointillist painting, displayed on the animal's skin. This process is orchestrated by many motor neurons, which innervate the radial muscles of individual chromatophores to control their contraction.

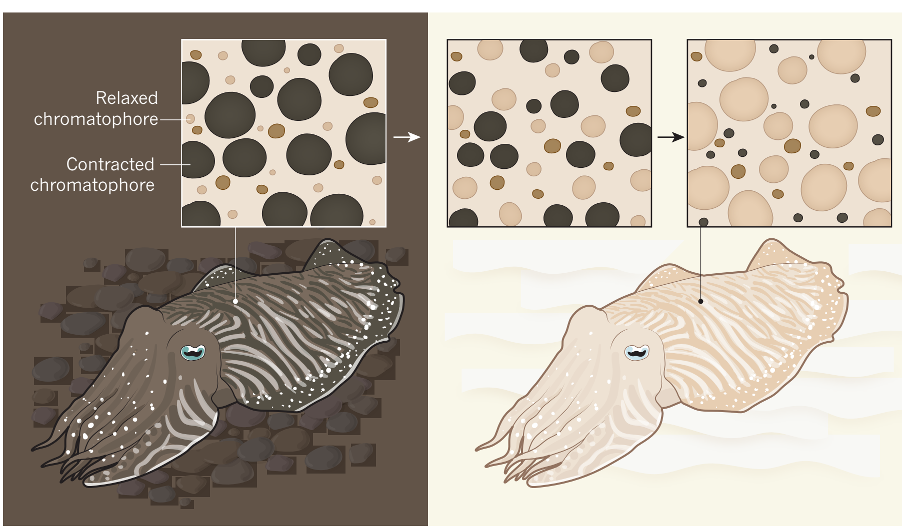


Figure 1 | Thoughts on display. Chromatophores are pigmented cells found on the skin of cuttlefish. Modulations in muscle contraction determine whether or not the cells' pigments are displayed, producing a changing patterning system that the animal uses for camouflage. Reiter *et al.*[1] used computer-vision tools to track tens of thousands of chromatophores. The authors' investigation reveals how skin pattern is controlled and how it varies over time. In response to changes in the cuttlefish's surroundings, the muscles that control groups of chromatophores contract or relax in unison, to produce a coordinated alteration in skin appearance.

Cuttlefish move rapidly, and because they are soft-bodied, frequently change shape. This constant flux presents a huge technical challenge for studies of individual chromatophores, because such analyses require imaging techniques that can keep track of individual cells between frames of video footage. Reiter *et al.* found that each chromatophore is surrounded by a unique arrangement of neighbouring chromatophores, akin to a fingerprint that could be picked out in a single frame, despite changes in skin pattern. By following the characteristic fingerprint of each chromatophore in the video footage, the researchers were able to simultaneously track tens of thousands of cells over time. This enabled them to study how the control of individual chromatophores produces the complex skin patterns formed by the cell population as a whole.

The authors first investigated the emergence of local skin motifs in which dark chromatophores are surrounded by more-colourful ones. Observations over several weeks led to a surprising discovery: the difference in colour reflects a difference in age. The pigment of every chromatophore starts as yellow before turning red, then brown, and ending up as black. New chromatophores are generated throughout the life of the cuttlefish, and the group found that the ratio of black to coloured chromatophores is maintained by keeping a tight balance between the birth rate of new cells and the time it takes them to mature to a black colour.

Reiter *et al.* showed that new chromatophores are generated in regions in which there are no existing ones, with a simple local-repulsion rule ensuring an even spread of the cells across the skin. The authors found that the same rule could explain the patterns of chromatophore formation seen in other species of cephalopod. These findings suggest that evolutionarily conserved molecular interactions govern chromatophore positioning — a proposal that should be investigated in the future. The rule also explains how the cuttlefish display system

maintains a steady functionality despite the animal's continuously increasing body size.

Next, the researchers investigated the dynamics of radial-muscle contraction and relaxation around tens of thousands of chromatophores. They discovered co-variations in muscle movements at many spatial scales, indicating that chromatophores are regulated by modules of motor neurons that function in synchrony, and that operate on skin patches of different sizes. The smallest modules consisted of fewer than ten adjacent chromatophores of the same colour. By contrast, larger modules, when contracted in synchrony, displayed more-complex shapes, such as rings, rectangles or disjointed structures resembling eye spots. These results pave the way to investigating how the geometry of these modules gives rise to the camouflage motifs seen in cuttlefish in their natural environment.

Finally, the authors studied chromatophore responses to changes in the cephalopod's visual environment, for instance when an investigator passed a hand above the animal, causing its skin pattern to change. They found that chromatophores display a highly coordinated choreography over time — reminiscent of the choreography of neuronal-population activity during movement[5]. Strikingly, chromatophores went through the same sequence of contractions and relaxations each time the test was repeated. This indicates a remarkable level of fine control by motor neurons, and highlights the potential of cuttlefish studies to deepen our understanding of complex motor systems.

Reiter *et al.* have achieved a breakthrough that will allow researchers to study this motor system in much more detail than was previously possible. The next challenge will be to determine how cuttlefish change the 3D texture of their skin for camouflage on sand, algae or corals. This process involves sets of muscles called papillae that create bumps and lumps. To gain a complete understanding of the animal's display system, chromatophores and papillae should be studied together.

The authors' advance also has implications for visual perception and motor control more generally. For instance, we should now be able to gain a better understanding of texture perception in both cephalopods and their vertebrate predators, by investigating which visual features in the cuttlefish environment drive skin-pattern choices. Given that we can read the perceptual state of cuttlefish on their skin, it might also become easier to investigate the brain activity that translates visual perceptions into motor outputs.

Furthermore, because cuttlefish coordinate millions of muscles simultaneously, they could provide insights into the principles underlying motor coordination. The authors' findings suggest a hierarchical organization of motor-neuron modules, in which higher-level modules control complex, global skin patterns and lower-level modules control simple, local motifs. Such a hierarchy of motor controllers has long been thought to be a key principle underlying behaviour in most animals, including humans[6]. However, recording the activity of every muscle in a human is currently impossible. The simple readout provided by the skin-display system of cuttlefish could well lead us to a greater understanding of motor control. ■

Adrien Jouary *and* **Christian K. Machens** *are in the Champalimaud Neuroscience Programme, Champalimaud Centre for the Unknown, 1400-038 Lisbon, Portugal.*
e-mails: adrien.jouary@neuro.fchampalimaud.org; christian.machens@neuro.fchampalimaud.org

1. Reiter, S. *et al. Nature* **562**, 361–366 (2018).
2. Mather, J. A. & Dickel, L. *Curr. Opin. Behav. Sci.* **16**, 131–137 (2017).
3. Hanlon, R. T. & Messenger, J. B. *Phil. Trans. R. Soc. B* **320**, 437–487 (1988).
4. Messenger, J. B. *Biol. Rev.* **76,** 473–528 (2001).
5. Churchland, M. M. *et al. Nature* **487**, 51–56 (2012).
6. Lashley, K. S. in *Cerebral Mechanisms in Behavior* (ed. Jefffries, L. A.) 112–136 (Wiley, 1951).

QUANTUM PHYSICS

Exploring the Universe with matter waves

An exotic ultracold gas known as a Bose–Einstein condensate has been produced and studied in space. Such gases could be used to build quantum sensors that probe the properties of the Universe with extreme precision. SEE LETTER P.391

LIANG LIU

Many great discoveries in modern physics depend on the invention of sensors based on new principles. For example, in 1887, an optical interferometer — a sensor based on wave interference — was used to disprove the existence of luminiferous aether, a universal medium through which light waves were thought to propagate[1]. In 1968, radio telescopes were used to discover extreme astronomical objects known as pulsars[2]. And in 2016, a laser interferometer was used to detect gravitational waves[3]. On page 391, Becker *et al.*[4] demonstrate how space-borne sensors based on an exotic state of matter called a Bose–Einstein condensate might provide the next big discovery.

A fundamental principle of quantum physics is wave–particle duality, which describes elementary particles in terms of quantum-mechanical waves (de Broglie

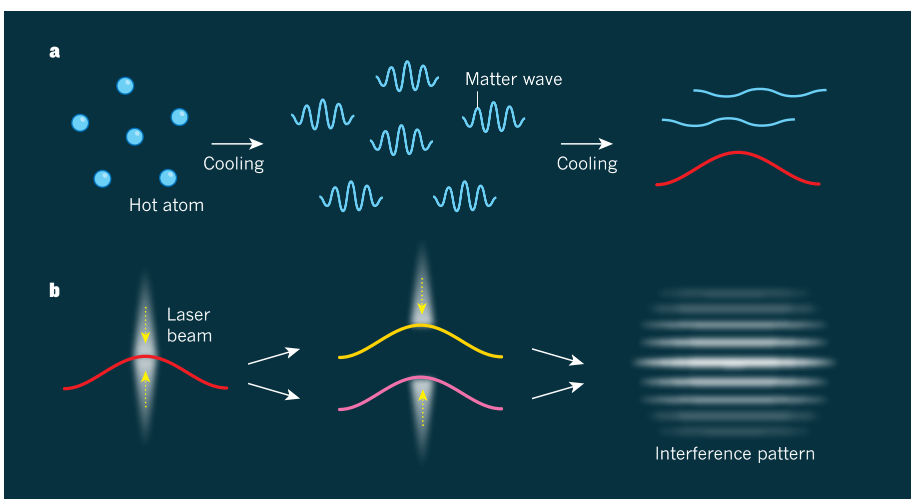


Figure 1 | Production and application of a Bose–Einstein condensate. **a**, In quantum physics, matter can behave like a wave that has a particular wavelength. For a cloud of hot atoms, these wavelengths are so short that each atom can be regarded as an individual object. If the atoms are cooled, the wavelengths become longer. And if the atoms are cooled to a critical temperature, the wavelengths are large enough to cover the extent of the atomic cloud. Most of the atoms condense into a state known as a Bose–Einstein condensate (BEC), in which they can be regarded as a single matter wave (red). Becker *et al.*[4] have produced and analysed a BEC in space. **b**, BECs can be used in sensors known as atom interferometers, in which laser beams cause a matter wave to split into two and then recombine to generate an interference pattern that is sensitive to external perturbations.

waves). The higher the velocity of a particle, the shorter the wavelength of the de Broglie wave. For a cloud of hot atoms, the de Broglie wavelengths are so short that each atom can be considered as an individual object (Fig. 1a).

If these atoms are cooled, the de Broglie wavelengths become longer. And if the atoms are cooled to a critical temperature (typically several hundred nanokelvin), the wavelengths become large enough to cover the whole atomic cloud. In this scenario, most of the atoms condense into a state in which they all behave in the same manner, and can be regarded as a single matter wave. Such a state is known as a Bose–Einstein condensate (BEC).

Producing a BEC is not easy. Even though the concept was proposed[5,6] in 1924–1925, a BEC was not realized[7,8] until 1995, after two types of cooling (laser and evaporative) had been invented. Since then, the matter waves associated with BECs have been widely used in atom interferometry (Fig. 1b). Atom interferometers use laser beams to split up matter waves and then recombine them to produce interference patterns. These patterns are sensitive to vibrations, changes in temperature and other disturbances.

Sensors based on matter waves differ from those based on light because atoms have a mass and an internal structure. The mass means that matter-wave sensors are extremely sensitive to gravity. They are therefore more suited to work in space, where gravity is extremely weak (a condition known as microgravity), than they are to work on the ground. Moreover, the internal structure of atoms means that there are more ways to control the properties of matter-wave sensors than those of optical sensors.

Becker and colleagues developed a BEC set-up for a rocket, which was launched to a height of 243 kilometres before returning to the ground. The BEC was produced while the rocket was in space, which is a milestone on the path towards building space-borne matter-wave sensors. During the launch phase and the 6 minutes of space flight, an astonishing 110 BEC-related experiments were carried out. The BEC set-up was only slightly bigger than the average human, withstood the vibrations and shocks during the launch of the rocket, and automatically conducted all of the experiments. Such a set-up represents a technical marvel in modern atomic physics.

The authors compared the formation of the BEC in space with that of one on the ground. They found that there were more atoms in the space-based BEC than in the ground-based one, although the fraction of atoms in the atomic cloud that were condensed was lower in space than on the ground. In an atom interferometer, a greater number of condensed atoms can give rise to a stronger interference signal, whereas a larger condensation fraction increases the signal-to-noise ratio. As a result, precision interferometry requires both a large number of condensed atoms and a high condensation fraction. The authors should therefore try to improve the condensation fraction for their space-borne BEC.

Becker *et al.* demonstrated transport of the BEC away from the surface of the chip on which it was formed — a key step towards realizing more-complex motion. Such motion, combined with further manipulation, would enable the natural expansion of the BEC to be precisely controlled, maximizing the time that the atomic cloud could be used in an interferometer. The transport of the BEC from the chip caused complex oscillations in the shape of the atomic cloud. These oscillations reveal valuable details about the hydrodynamic behaviour of the BEC, but their impact on interferometry performance needs further investigation.

On the ground, microgravity can be achieved for only a few seconds. But in space, it can be supported for essentially an infinite length of time, offering new opportunities for studying cold-atom physics. For example, a BEC in microgravity could reach temperatures as low as picokelvin (equal to 10^{-12} K) or even femtokelvin (10^{-15} K) ranges, compared with nanokelvin on the ground. Gases at such low temperatures are an ideal platform for probing fundamental physics, and the authors' space-borne BEC is the first step towards this goal.

Becker and colleagues' work paves the way for quantum sensors in space that could be used to conduct experiments that are not possible on Earth. Examples include detecting gravitational waves in a frequency range that is not usually accessible, sensing possible ultralight dark-matter particles and observing subtle effects associated with Einstein's general theory of relativity. Who knows what mysteries of the Universe could be revealed by space-borne quantum sensors. ■

Liang Liu *is in the Key Laboratory of Quantum Optics, Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai 201800, China.*
e-mail: liang.liu@siom.ac.cn

1. Michelson, A. A. & Morley, E. W. *Am. J. Sci.* **34**, 333–345 (1887).
2. Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F. & Collins, R. A. *Nature* **217**, 709–713 (1968).
3. Abbott, B. P. *et al. Phys. Rev. Lett.* **116**, 061102 (2016).
4. Becker, D. *et al. Nature* **562**, 391–395 (2018).
5. Bose, S. N. *Z. Phys.* **26**, 178–181 (1924).
6. Einstein, A. *Phys. Math. Klasse* **1**, 3–14 (1925).
7. Anderson, M. H., Ensher, J. R., Matthews, M. R., Wieman, C. E. & Cornell, E. A. *Science* **269**, 198–201 (1995).
8. Davis, K. B. *et al. Phys. Rev. Lett.* **75**, 3969–3974 (1995).

CELLULAR EVOLUTION

The eukaryotic ancestor shapes up

Asgard archaea are the closest known relatives of nucleus-bearing organisms called eukaryotes. A study indicates that these archaea have a dynamic network of actin protein — a trait thought of as eukaryote-specific. **See Letter p.439**

LAURA EME & THIJS J. G. ETTEMA

Eukaryotic cells, which carry their DNA in a nucleus, are thought to have evolved from a merger between two other organisms — an archaeal host cell[1–3] and a bacterium from which eukaryotic organelles called mitochondria emerged[4]. Some insights into the biological properties of the host have come from the closest known archaeal relatives of eukaryotes, the Asgard superphylum[5,6]. The genomes of organisms belonging to this archaeal group encode a suite of proteins typically involved in functions or processes thought to be eukaryote-specific. The functions of these 'eukaryotic genes' in Asgard archaea have been elusive, but on page 439, Akıl and Robinson[7] provide evidence that some of them encode proteins that are structurally and functionally similar to their eukaryotic counterparts.

Apart from their nucleus and energy-producing mitochondria, eukaryotic cells are characterized by a complex internal system of membrane-bound compartments (the endomembrane system), and by a dynamic network of proteins such as actin, called the cytoskeleton. The latter gives the cells their shape and structure, but is also involved in a variety of cellular processes specific to eukaryotes[8]. These features are thought to have been present in the last common ancestor of all eukaryotes, which lived about 1.8 billion years ago[9], but no life forms have been found that represent an intermediate between eukaryotes and their bacterial and archaeal ancestors. The seemingly sudden emergence of cellular complexity in the eukaryotic lineage is a conundrum for evolutionary biologists.

Several of the proteins produced by Asgard archaea are evolutionarily related to proteins that in eukaryotes modulate complex cellular processes[5,6]. The identification of these proteins raised the question of whether Asgard archaea have some primitive versions of

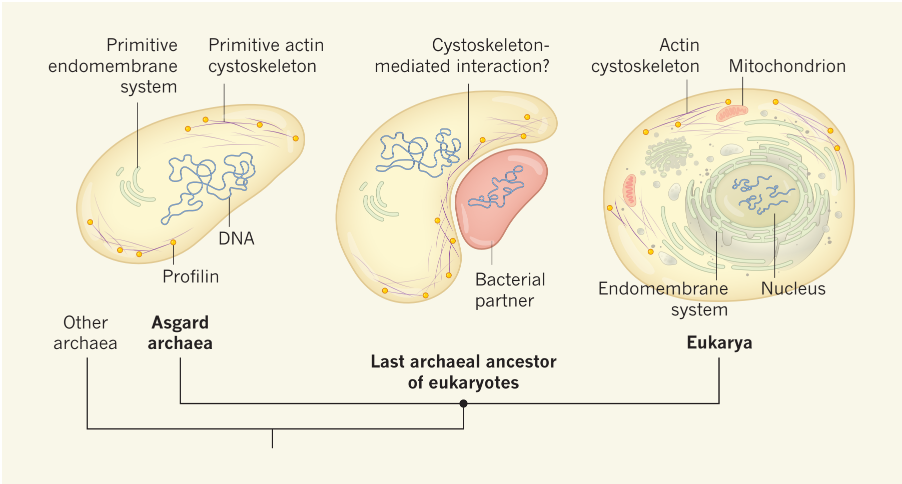


Figure 1 | Cellular complexity along the tree of life. The Eukarya (organisms whose cells harbour DNA in a nucleus) are thought to have arisen from a merger between their last archaeal ancestor and a bacterium. In addition to a nucleus, eukaryotes have several characteristics that are thought to separate them from archaea, including: a complex internal system of membranes called endomembranes; a structural feature called the actin cytoskeleton, the dynamics of which are regulated by the protein profilin; and energy-producing organelles called mitochondria, which arose from the bacterial partner. But Akıl and Robinson[7] provide evidence that members of the Asgard superphylum — an extant group of archaea thought to be related to eukaryotes — harbour a primitive profilin-regulated actin cytoskeleton. If the last archaeal ancestor of eukaryotes had this feature, it might have enabled the cell to wrap around its presumed bacterial partner. In addition, it is possible that Asgard archaea and the last archaeal ancestor of eukaryotes carry primitive endomembrane systems. (Cells and cellular features are not drawn to scale.)

certain eukaryotic properties. If they do, it would suggest that the last archaeal ancestor of eukaryotes already displayed a certain — albeit probably limited — degree of cellular complexity reminiscent of eukaryotes.

Experiments to support such ideas are complicated by the fact that evidence for the existence of the four known Asgard lineages (Lokiarchaeota, Odinarchaeota, Thorarchaeota and Heimdallarchaeota)[5,6] is based solely on metagenomics analyses. The cells have yet to be observed under a microscope, and have not been cultured *in vitro*. Nevertheless, Akıl and Robinson were determined to gain insight into the properties of Asgard proteins related to the eukaryotic proteins actin and profilin. In eukaryotes, profilin regulates the polymerization of actin into filaments of the cytoskeleton. These filaments have pivotal roles in processes that include vesicle and organelle movement, cell-shape formation and phagocytosis[8], in which cells ingest foreign particles or other cells.

To produce Asgard profilins, Akıl and Robinson expressed these proteins in the bacterium *Escherichia coli* using a circular DNA molecule called a plasmid that harboured the profilin-encoding genes. They then purified the proteins and studied their structures using X-ray crystallography. Asgard profilins share limited amino-acid sequence identity with their eukaryotic counterparts. Nonetheless, the authors found that the structure of loki-archaeal profilin is topologically similar to that of human profilin, although some structural divergences could be observed. This confirms that Asgard and eukaryotic profilins are indeed evolutionarily related, albeit distantly.

Next, the researchers set out to investigate whether Asgard profilins could interact with Asgard actins. Unfortunately, despite considerable efforts, they were unable to produce functional Asgard actin. As an alternative, they therefore carried out *in vitro* and co-crystallization experiments to test whether Asgard profilins could interact with eukaryotic actins. Remarkably, despite being separated by 2 billion to 3 billion years of evolution[9], several of the Asgard profilins bound to mammalian actin and regulated its polymerization kinetics. Asgard and mammalian profilins seem to have similar effects on mammalian actin, although the Asgard proteins act less efficiently. These results suggest that Asgard archaea harbour a profilin-regulated actin cytoskeleton — a cellular feature generally regarded as a defining characteristic of eukaryotic cells (Fig. 1).

"The inference of a primitive dynamic actin cytoskeleton in Asgard archaea sheds light on the biological properties of the ancestor of eukaryotes."

The inference of a primitive dynamic actin cytoskeleton in Asgard archaea sheds light on the biological properties of the ancestor of eukaryotes. In eukaryotic cells, the energy required to dynamically regulate actin is mainly provided by mitochondria[10]. Although the energetic and metabolic properties of Asgard archaea are currently unknown, they certainly lack the firepower that mitochondria provide. A profilin-regulated actin cytoskeleton in the archaeal ancestor of eukaryotes is therefore unlikely to sustain energy-consuming processes such as phagocytosis.

But was the energy provided by mitochondria necessarily the ultimate driving force for the emergence of complex cellular features in eukaryotes? Archaea such as *Ignicoccus hospitalis*, along with several types of bacterium, have independently evolved endomembrane systems[11]. Because these lineages lack mitochondria, energetic constraints can be ruled out as a limiting factor in the emergence of such a system. It is therefore feasible that Asgard archaeal cells produce sufficient energy to harbour both a primitive endomembrane system and undergo actin-driven membrane and cell-shape deformation. Perhaps the latter ability could have facilitated the symbiotic interaction between the Asgard-related host cell and the bacterial ancestor of mitochondria, for example by optimizing the membrane surface area for metabolic exchange between the two cells. Once mitochondria became an intrinsic part of eukaryotic cells, their capacity for energy production could have conferred selective advantages on their host. However, the exact contribution of these organelles to the emergence of the complex features of eukaryotic cells remains unresolved.

Future efforts to elucidate the biological and physiological properties of Asgard archaea will be essential to increase our understanding of the emergence of eukaryotes. Although biochemical and structural studies of individual Asgard proteins, such as those by Akıl and Robinson, are likely to provide piecemeal insights, it is the ability to grow Asgard archaeal lineages *in vitro* that will ultimately unravel their obscure biology. ■

Laura Eme *and* **Thijs J. G. Ettema** *are in the Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, 75123 Uppsala, Sweden.*
e-mail: thijs.ettema@icm.uu.se

1. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. *Nature* **504**, 231–236 (2013).
2. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. *Proc. Natl Acad. Sci. USA* **112**, 6670–6675 (2015).
3. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. *Nature Rev. Microbiol.* **15**, 711–723 (2017).
4. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. *Curr. Biol.* **27**, R1177–R1192 (2017).
5. Spang, A. *et al. Nature* **521**, 173–179 (2015).
6. Zaremba-Niedzwiedzka, K. *et al. Nature* **541**, 353–358 (2017).
7. Akıl, C. & Robinson, R. C. *Nature* **562**, 439–443 (2018).
8. Blanchoin, L., Boujemaa-Paterski, R., Sykes, C. & Plastino, J. *Physiol. Rev.* **94**, 235–263 (2014).
9. Betts, H. C. *et al. Nature Ecol. Evol.* https://doi.org/10.1038/s41559-018-0644-x (2018).
10. Martin, W. F., Tielens, A. G. M., Mentel, M., Garg, S. G. & Gould, S. B. *Microbiol. Mol. Biol. Rev.* **81**, e00008–17 (2017).
11. Grant, C. R., Wan, J. & Komeili, A. *Annu. Rev. Cell Dev. Biol.* https://doi.org/10.1146/annurev-cellbio-100616-060908 (2017).

This article was published online on 3 October 2018.

https://doi.org/10.1038/s41586-018-0599-8

Improved limit on the electric dipole moment of the electron

ACME Collaboration*

The standard model of particle physics accurately describes all particle physics measurements made so far in the laboratory. However, it is unable to answer many questions that arise from cosmological observations, such as the nature of dark matter and why matter dominates over antimatter throughout the Universe. Theories that contain particles and interactions beyond the standard model, such as models that incorporate supersymmetry, may explain these phenomena. Such particles appear in the vacuum and interact with common particles to modify their properties. For example, the existence of very massive particles whose interactions violate time-reversal symmetry, which could explain the cosmological matter–antimatter asymmetry, can give rise to an electric dipole moment along the spin axis of the electron. No electric dipole moments of fundamental particles have been observed. However, dipole moments only slightly smaller than the current experimental bounds have been predicted to arise from particles more massive than any known to exist. Here we present an improved experimental limit on the electric dipole moment of the electron, obtained by measuring the electron spin precession in a superposition of quantum states of electrons subjected to a huge intramolecular electric field. The sensitivity of our measurement is more than one order of magnitude better than any previous measurement. This result implies that a broad class of conjectured particles, if they exist and time-reversal symmetry is maximally violated, have masses that greatly exceed what can be measured directly at the Large Hadron Collider.

The electric dipole moment (EDM) of the electron is an asymmetric charge distribution along the particle's spin. The existence of an EDM requires violation of time-reversal symmetry. The standard model of particle physics predicts that the electron has such an EDM, d_e, but with a magnitude far below current experimental sensitivities[1–3]. However, theories of physics beyond the standard model generally include new particles and interactions that can break time-reversal symmetry. If these new particles have masses of 1–100 TeV c^{-2}, theories typically predict that $d_e \approx 10^{-27}$–$10^{-30} e$ cm ($1e$ cm $= 1.6 \times 10^{-21}$ C m, where e is the electron charge)[4–8]—a value that is orders of magnitude larger than the standard model predictions, which is now accessible by experiment[1,9]. Here we report the result of the ACME II experiment, an improved measurement of d_e with sensitivity over 10 times better than the previous best measurement, ACME I[1,9]. This was achieved by improving the state preparation, experimental geometry, fluorescence collection and control of systematic uncertainties. Our measurement, $d_e = (4.3 \pm 3.1_{\rm stat} \pm 2.6_{\rm syst}) \times 10^{-30} e$ cm ('stat', statistical uncertainty; 'syst', systematic uncertainty), is consistent with zero and corresponds to an upper limit of $|d_e| < 1.1 \times 10^{-29} e$ cm at 90% confidence. This result constrains new time-reversal-symmetry-violating physics for broad classes of proposed beyond-standard-model particles with masses in the range 3–30 TeV c^{-2}.

Recent advances in the measurement of d_e[1,10–12] have relied on using the exceptionally high internal effective electric field ($\mathcal{E}_{\rm eff}$) of heavy polar molecules[13–15]. This gives rise to an energy shift $U = -\boldsymbol{d}_e \cdot \boldsymbol{\mathcal{E}}_{\rm eff}$, where $\boldsymbol{d}_e = d_e \boldsymbol{s}/(\hbar/2)$, $\boldsymbol{s}$ is the spin of the electron and $\hbar$ is the reduced Planck constant. The $H^3\Delta_1$ electronic state in the thorium monoxide (ThO) molecule has[16,17] $\mathcal{E}_{\rm eff} \approx 78$ GV cm^{-1} when the molecule is fully polarized; this requires only a very modest electric field ($\mathcal{E} \gtrsim 1$ V cm^{-1}) applied in the laboratory. ACME I used ThO to place a limit of $|d_e| < 9.4 \times 10^{-29} e$ cm (90% confidence)[1,9], which was recently confirmed by an experiment with trapped HfF^+ molecular ions[12], which found $|d_e| < 1.3 \times 10^{-28} e$ cm.

An EDM measurement with thorium monoxide

As in ACME I, we performed our measurement in the $J = 1$, $M = \pm 1$ sublevels of the $H^3\Delta_1$ state of ThO, where J is the angular momentum and M is its projection along a quantization axis $\hat{\boldsymbol{z}}$ (Fig. 1a). In our applied electric field $\boldsymbol{\mathcal{E}} = \mathcal{E}_z \hat{\boldsymbol{z}}$, these states are fully polarized[18], such that the internuclear axis $\hat{\boldsymbol{n}}$, which points from the oxygen to the thorium nucleus, is either aligned or antialigned with $\boldsymbol{\mathcal{E}}$. The direction of $\hat{\boldsymbol{n}}$ coincides with the direction of the field $\boldsymbol{\mathcal{E}}_{\rm eff}$ that acts on $\boldsymbol{d}_e$. States with opposite molecule orientation are described by the quantum number $\tilde{\mathcal{N}} = {\rm sgn}(\boldsymbol{\mathcal{E}} \cdot \hat{\boldsymbol{n}}) = \pm 1$. The direction of $\boldsymbol{\mathcal{E}}_{\rm eff}$ can be reversed either by reversing the laboratory field $\boldsymbol{\mathcal{E}}$ or by changing the state $\tilde{\mathcal{N}} = \pm 1$ used in the measurement; each of these approaches allows us to reject a wide range of systematic errors[19–21].

The electron spin, $\boldsymbol{s}$, is along the spin of the molecular state, $\boldsymbol{S}$. We measure the energy difference between states with $M = \pm 1$ (which correspond to $\boldsymbol{S}$ being aligned or antialigned with $\boldsymbol{\mathcal{E}}_{\rm eff}$; Fig. 1a), which contains a term proportional to U. To do so, we prepare an initial coherent superposition of $M = \pm 1$ states, which corresponds to the spin $\boldsymbol{S}$ being aligned with a fixed direction in the x–y plane (Fig. 2). The applied magnetic field, $\boldsymbol{\mathcal{B}} = \mathcal{B}_z \hat{\boldsymbol{z}}$, and $\boldsymbol{\mathcal{E}}_{\rm eff}$ exert torques on the magnetic and electric dipole moments associated with the spin, causing $\boldsymbol{S}$ to precess in the x–y plane by an angle ϕ as the molecules travel freely. The final value of ϕ is measured by laser excitation of the molecules, which induces fluorescence with a strength that depends on the angle between $\boldsymbol{S}$ and the laser polarization. The angle ϕ is given by

$$\phi \approx \frac{-(\mu \tilde{\mathcal{B}} |\mathcal{B}_z| + \tilde{\mathcal{N}} \tilde{\mathcal{E}} d_e \mathcal{E}_{\rm eff})\tau}{\hbar} \quad (1)$$

where $|\mathcal{B}_z| = |\boldsymbol{\mathcal{B}} \cdot \hat{\boldsymbol{z}}|$, $\tilde{\mathcal{B}} = {\rm sgn}(\boldsymbol{\mathcal{B}} \cdot \hat{\boldsymbol{z}})$, $\tilde{\mathcal{E}} = {\rm sgn}(\boldsymbol{\mathcal{E}} \cdot \hat{\boldsymbol{z}})$, τ is the spin precession time and $\mu = \mu_B g_{\mathcal{N}}$, where $g_{\mathcal{N}} = -0.0044$ is the g-factor of the $|{\rm H}, J = 1, \mathcal{N}\rangle$ state[22] and μ_B is the Bohr magneton. The sign, $\tilde{\mathcal{N}}\tilde{\mathcal{E}}$, of the EDM contribution to the angle is given by the sign of the torque of $\boldsymbol{\mathcal{E}}_{\rm eff}$

*A list of participants and their affiliations appears at the end of the paper.

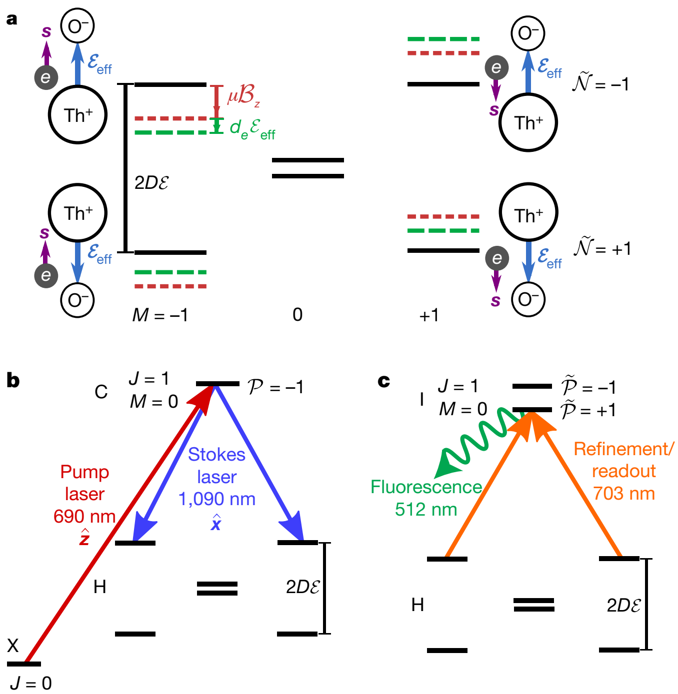


Fig. 1 | Energy levels of thorium monoxide and laser transitions. The addressed transitions are shown for one of several possible experimental states. **a**, Levels of the state H, $J = 1$ in external electric ($\boldsymbol{\mathcal{E}}$) and magnetic ($\boldsymbol{\mathcal{B}}$) fields. The orientation of the effective electric field, $\boldsymbol{\mathcal{E}}_{\mathrm{eff}}$, is shown by blue arrows and that of the spin of the electron, $\boldsymbol{s}$, by purple arrows. The energy shifts $\mu\mathcal{B}_z$ (brown) and $d_e\mathcal{E}_{\mathrm{eff}}$ (green) due to the magnetic moment μ and the EDM d_e, respectively, are shown. The $\tilde{\mathcal{N}} = \pm 1$ states are split by $2D\mathcal{E} \approx 200$ MHz owing to the Stark effect, where D is the H-state electric dipole moment. **b**, STIRAP efficiently transfers population from the ground state $|X, J=0\rangle$ to a spin-aligned superposition of one molecule orientation, $\tilde{\mathcal{N}} = +1$ or $\tilde{\mathcal{N}} = -1$ ($\tilde{\mathcal{N}} = -1$ shown here). STIRAP uses two lasers, the pump laser (red arrow; X–C, 690 nm, polarized along $\hat{z}$) and the Stokes laser (blue arrows; C–H, 1,090 nm, polarized along $\hat{x}$). **c**, The refinement laser (orange) removes imperfections in the spin-aligned state prepared by STIRAP. The readout laser (orange) excites the molecule from its original orientation, $\tilde{\mathcal{N}} = +1$ or $\tilde{\mathcal{N}} = -1$ ($\tilde{\mathcal{N}} = -1$ shown here), to an isolated $J = 1$, $M = 0$ level in state I. This state can have either parity, $\tilde{\mathcal{P}} = +1$ or $\tilde{\mathcal{P}} = -1$ ($\tilde{\mathcal{P}} = +1$ shown here). The I state decays via spontaneous photon emission, and we detect the resulting fluorescence (green wavy arrow).

on $\boldsymbol{s}$. The spin precession frequency, $\omega = \phi/\tau$, is given by the energy shift between the $M = \pm 1$ states (divided by $\hbar$). The value of d_e is extracted from the change in ω that is correlated with the orientation of $\boldsymbol{\mathcal{E}}_{\mathrm{eff}}$ in the laboratory frame, that is, with the product $\tilde{\mathcal{N}}\tilde{\mathcal{E}}$. By denoting this correlated component as $\omega^{\mathcal{N}\mathcal{E}}$, we obtain $d_e = -\hbar\omega^{\mathcal{N}\mathcal{E}}/\mathcal{E}_{\mathrm{eff}}$.

We produce ThO molecules in a cryogenic buffer gas beam source[23–25]. The molecules pass through laser beams and are rotationally cooled, increasing the population of the lowest energy level (ground electronic state X, rotational level $J = 0$) by a factor of 2.5. The ThO molecules then enter a magnetically shielded region where the EDM measurement is performed. The electric field $\boldsymbol{\mathcal{E}}$ is produced by a set of parallel plates and the magnetic field $\boldsymbol{\mathcal{B}}$ is generated by a current circulating through coils (Fig. 2). We prepare the desired initial spin state using stimulated Raman adiabatic passage (STIRAP), coherently transferring the molecules from the ground state $|X, J=0\rangle$ to a specific sublevel of the lowest rotational level, $J = 1$, of the metastable (lifetime 2 ms)[18] electronic $H^3\Delta_1$ state manifold[26] (Fig. 1b). This results in a coherent superposition of $M = \pm 1$ states. STIRAP is implemented through a pair of co-propagating laser beams (wavelengths of 690 nm and 1,090 nm) resonant with the electronic transitions X–C and C–H. These beams are partially spatially overlapped, travel vertically (along $\hat{y}$) and have linear polarizations along $\hat{z}$ and $\hat{x}$, respectively. We choose which $\tilde{\mathcal{N}}$ state to address by tuning the frequency of the H–C STIRAP laser. The technical details of the STIRAP implementation are given in a separate publication[26].

Imperfections in the STIRAP-prepared spin-aligned state can lead to systematic errors but are suppressed with the following method. After leaving the STIRAP region, the molecules enter a linearly polarized 'refinement' laser that optically pumps away the unwanted spin component and leaves behind a dark superposition of the two resonant $M = \pm 1$ sublevels[27] of H. The refinement laser is resonant with the H–I transition (wavelength 703 nm; Fig. 1c). Within the short-lived (lifetime 115 ns) electronic state I, there are two well resolved opposite-parity ($\tilde{\mathcal{P}} = \pm 1$) states with $J = 1$ and $M = 0$[28,29]. The refinement laser polarization is nominally aligned with the STIRAP-prepared spin $\boldsymbol{S}_{\mathrm{ST}}$ and addresses the $\tilde{\mathcal{P}} = +1$ parity state in I. The resulting refined state, $|\psi(t=0), \tilde{\mathcal{N}}\rangle$, has $\boldsymbol{S}$ aligned with $\hat{\boldsymbol{x}}$ more accurately than the initial STIRAP-prepared state (Fig. 2).

Molecules travel over a distance of $L \approx 20$ cm (corresponding to $\tau \approx 1$ ms) so that $\boldsymbol{S}$ precesses in the x–y plane by angle ϕ (given by equation (1)). This yields the molecular state at time $t = \tau$,

$$|\psi(t=\tau), \tilde{\mathcal{N}}\rangle = \frac{e^{-i\phi}|M=+1, \tilde{\mathcal{N}}\rangle - e^{+i\phi}|M=-1, \tilde{\mathcal{N}}\rangle}{\sqrt{2}} \quad (2)$$

We measure ϕ by exciting the H–I transition with laser light linearly polarized along direction $\hat{\epsilon}$. This yields fluorescence signals with intensity S_ϵ, which depends on the angle between $\hat{\epsilon}$ and $\boldsymbol{S}$. To remove the effects of fluctuations in molecule number, we excite the molecules with two alternating orthogonal linear polarizations, $\hat{\epsilon} = \hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}}$, by modulating $\hat{\epsilon}$ sufficiently rapidly (period 5 μs) so that each molecule is addressed by both polarizations as it flies through the laser beam[22]. We record the corresponding fluorescence signals S_X and S_Y from the decay of I to the ground state X (wavelength 512 nm; see Extended Data Fig. 1a). We then compute the asymmetry[18]

$$\mathcal{A} = \frac{S_X - S_Y}{S_X + S_Y} = \mathcal{C}\cos[2(\phi - \theta)] \quad (3)$$

where the contrast $\mathcal{C}$ is 95% ± 2% on average and $\hat{\boldsymbol{X}}$ is defined to be at an angle θ with respect to $\hat{\boldsymbol{x}}$ in the x–y plane (Fig. 2). This procedure amounts to a projective measurement of the molecule alignment onto both $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{Y}}$. We set $|\mathcal{B}_z|$ and θ such that $\phi - \theta \approx (\pi/4)(2n+1)$ for integer n, so that the asymmetry is linearly proportional to small changes in ϕ and thus maximally sensitive to d_e. We measure $\mathcal{C}$ by dithering θ between two nearby values, $\tilde{\theta} = \pm 1$, that differ by 0.2 rad.

When limited by shot noise, the uncertainty in the measured phase, $\delta\phi$, per unit of measurement time scales as the square root of the photoelectron detection rate[22]. Compared with ACME I, ACME II improves phase sensitivity by an order of magnitude by increasing the fraction of beam source molecules used in the measurement. The implementation of STIRAP, together with a redesigned rotational-cooling scheme, improves the state preparation efficiency by a factor of 12. The detected solid angle of the diverging molecular beam is increased by a factor of 7 by moving the source closer to the detection region and increasing the separation between the electric-field plates, the size of all laser beams and the openings of the molecular beam collimators. The photon collection efficiency is increased by a factor of 5 using a combination of detecting shorter-wavelength photons (512 nm in ACME II, compared with 690 nm in ACME I), for which the photomultiplier tubes (PMTs) have higher quantum efficiency, and by replacing fibre bundles with lower-loss solid glass light pipes to transfer light to the PMTs. Together, these improvements increase our photoelectron detection rate by a factor of about 400 over ACME I.

We performed repeated spin precession measurements under varying experimental conditions to (a) isolate the EDM phase from background phases and (b) search for and monitor possible systematic errors. Within a 'block' of data (Extended Data Fig. 1c) taken over 60 s, we performed four identical measurements of ϕ for each state in the

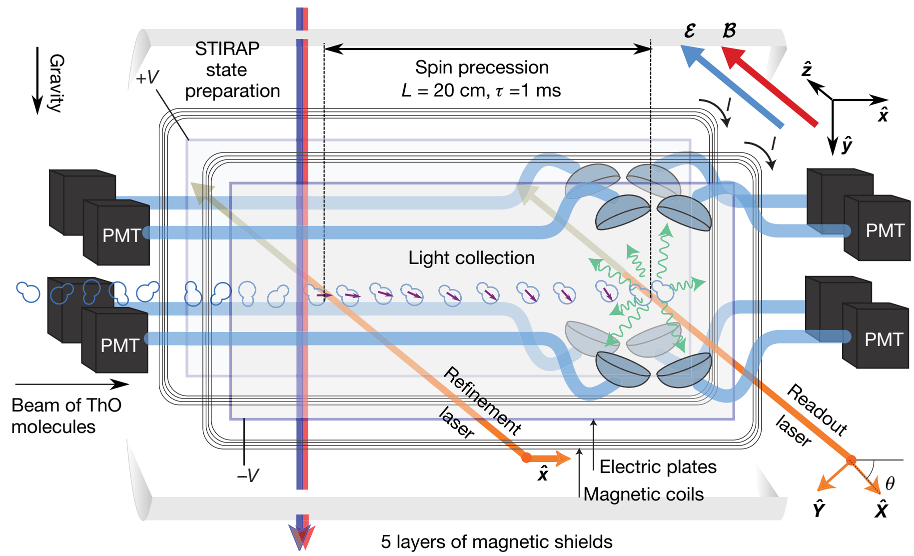


Fig. 2 | Schematic of the measurement region. A collimated pulse of ThO molecules enters a magnetically shielded region. A uniform electric field $\boldsymbol{\mathcal{E}}$ is applied by a set of transparent parallel plates at voltage $(+V, -V)$, and a uniform magnetic field $\boldsymbol{\mathcal{B}}$ is applied by circulating a current I through coils. A spin state (purple arrows) aligned along $\hat{\boldsymbol{x}}$, prepared by STIRAP (blue and red vertical arrows) and refined via an optical pumping laser beam polarized along $\hat{\boldsymbol{x}}$ (left orange arrow), precesses over a length of $L \approx 20$ cm (time $\tau \approx 1$ ms) in the applied electric and magnetic fields. The final spin alignment direction is read out by a laser (right orange arrow) with rapidly alternating linear polarizations, $\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}}$ (with $\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{X}}$ at an angle θ with respect to $\hat{\boldsymbol{x}}$). The resulting fluorescence (green wavy arrows), the intensity of which depends on the angle between the spin of the molecular state, $\boldsymbol{S}$, and $\hat{\boldsymbol{\epsilon}}$, is collected and detected using photomultiplier tubes (PMTs).

complete set of 2^4 experimental states derived from four binary switches: the molecular alignment, $\tilde{\mathcal{N}}$; the direction of the applied electric field, $\tilde{\mathcal{E}}$; the readout-laser polarization dither state, $\tilde{\theta}$; and the magnetic field direction, $\tilde{\mathcal{B}}$. We form 'switch-parity components' of the phase, which are combinations of the measured phases that are odd or even under the selected switch operations[21]. We denote the experimental parity of a quantity with a superscript, u, that lists all the switches under which the quantity has odd parity (the parity of the quantity is even for all switches not included in the superscript), and we use the superscript 'nr' to indicate that the quantity is even under all considered switches. For example, we extract d_e from the $\phi^{\mathcal{N}\mathcal{E}}$ component of the phase (see equation (1)), which is odd under the $\tilde{\mathcal{N}}$ and $\tilde{\mathcal{E}}$ switches and even under all other switches. We extract the precession time τ from the component of the phase that is odd under only the $\tilde{\mathcal{B}}$ switch, $\phi^{\mathcal{B}} = -\mu|\mathcal{B}_z|\tau/\hbar$, and use it to compute the frequency components $\omega^u = \phi^u/\tau$ that are odd under the chosen parity u.

On a slower timescale, we perform additional 'superblock' binary switches to suppress known systematic errors and to search for unknown ones (Extended Data Fig. 1d). These switches are: $\tilde{\mathcal{P}}$, the excited-state parity addressed by the readout laser; $\tilde{\mathcal{L}}$, the interchange of the supplies that apply voltage to the two electric-field plates; and $\tilde{\mathcal{R}}$, the rotation of the readout $\hat{\boldsymbol{X}}$–$\hat{\boldsymbol{Y}}$ polarization basis by $\pi/2$, $\theta \to \theta + \pi/2$. The $\tilde{\mathcal{P}}$ and $\tilde{\mathcal{R}}$ switches both interchange the role of $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{Y}}$ and thus reject systematic errors associated with small changes in the power, profile or pointing of the readout laser beam when the polarization $\hat{\boldsymbol{\epsilon}}$ is changed. The $\tilde{\mathcal{L}}$ switch rejects systematic errors that are proportional to the offset voltage of the electric-field power supplies. To compute d_e, we extract from the 2^7 block and superblock states $\omega^{\mathcal{N}\mathcal{E}}$, the component of the frequency that is odd under $\tilde{\mathcal{N}}$ and $\tilde{\mathcal{E}}$ and even under all other switches.

The EDM dataset consists of about 20,000 blocks, taken over the course of about two months (Extended Data Fig. 1f). During the acquisition of this dataset, in addition to the 7 switches described above, we also varied the magnitude of the magnetic field as $|\mathcal{B}_z| = 0.7$ mG, 1.3 mG, 2.6 mG and 26 mG (corresponding to $|\phi| \approx \pi/160$, $2\pi/160$, $4\pi/160$ and $\pi/4$, respectively), and the magnitude of the electric field as $|\mathcal{E}_z| = 80$ V cm^{-1} and 140 V cm^{-1}. 5% of the data were taken with $|\mathcal{B}_z| = 2.6$ mG; the rest were taken at $|\mathcal{B}_z| = 0.7$ mG, 1.3 mG and 26 mG in approximately equal amounts. Equal amounts of data were taken with each of the two electric field magnitudes. The $\omega^{\mathcal{N}\mathcal{E}}$ values obtained by isolating the data under each of these parameter values are shown in Fig. 3c.

Statistics of the EDM dataset

During data acquisition, we average 25 molecular pulses together to form a 'trace' (Extended Data Fig. 1b) and record individual traces corresponding to each of the eight PMTs. We typically sum the photoelectron signals in the eight PMTs but also frequently check the spatial dependence of the fluorescence as a diagnostic. Within a trace, we compute $\mathcal{A}$ for each polarization cycle (Extended Data Fig. 1a). We then average 20 cycles into a single 'group', with the uncertainty defined as the standard error in the mean of the group. The uncertainties of all groups are consistent with the level of shot noise in our photoelectron signals. We then use standard uncertainty propagation methods to compute the uncertainties from an entire superblock.

The scatter in the superblock data is found to be larger than that expected from group-level uncertainties. This noise is present in all switch-parity components. The excess noise in the precession frequency has one contribution that is proportional to the magnitude of the magnetic field and another that is independent of it; the latter component results in a reduced χ^2 of $\chi_r^2 \approx 3$. Because our fastest switch, $\tilde{\mathcal{N}}$, does not remove such noise, it enters the measurement at timescales lower than 0.6 s. We observed an increase in this noise after switching to a different ablation laser. This suggests that this noise might be related to fluctuations in the ablation characteristics, which are known to be correlated with molecular beam properties such as flux and transverse velocity.

The second component of the excess noise increases the scatter of our superblock data to $\chi_r^2 \approx 7$, but only for the largest applied magnetic field, $|\mathcal{B}_z| = 26$ mG. We verified through simulations and a direct measurement that this is consistent with about 0.05% shot-to-shot fluctuations in the mean longitudinal molecular velocity ($\langle v \rangle \approx 200$ m s^{-1}). Because the refinement and readout beams are fixed in space, variations in $\langle v \rangle$ change the precession time τ; which causes variations in the phase ϕ, which is proportional to $|\mathcal{B}_z|$ (for $d_e = 0$), as shown in equation (1). To reduce its effect, we acquired most of the data at lower magnetic fields, where the associated increase in χ_r^2 is negligible.

To prevent experimenter bias, we performed a blind analysis by adding an unknown offset to $\omega^{\mathcal{N}\mathcal{E}}$. We revealed this offset only after the

data collection, data cuts and two independent error analyses were complete. Figure 3a, b shows the distribution of the $\omega^{\mathcal{NE}}$ superblock data. The majority of the data are consistent with a Gaussian distribution, but with more points in the tails. We performed a robust M-estimator analysis[30] on bootstrapped[31,32] sets of data to extract confidence intervals corresponding to 1σ (68% confidence). Because the noise that arises from fluctuations in the mean longitudinal velocity depends on $|\mathcal{B}_z|$, we performed separate M-estimator analyses on subsets of data with different $|\mathcal{B}_z|$ values and then combined the sets using standard uncertainty propagation methods.

Systematic error investigations

To search for possible sources of systematic error, we varied over 40 different experimental parameters over larger ranges than those typically used in the experiment (Extended Data Table 1) and measured their effect on $\omega^{\mathcal{NE}}$ and the other parity components of both ω and $\mathcal{C}$. In particular, we varied a parameter P over a range ΔP and, by assuming a linear relation between P and $\omega^{\mathcal{NE}}$, determined the slope $S_P = \partial\omega^{\mathcal{NE}}/\partial P$. Such data, taken with intentionally applied parameter imperfections (that is, P set to a non-zero value although its ideal value is zero), were used only for the determination of systematic shifts and uncertainties, and were not included in the EDM dataset.

We used these measured slopes to compute systematic shifts and uncertainties as follows. If S_P was either expected or observed to be non-zero, we measured the residual ambient deviation of P from its ideal value, dP, and computed an associated systematic shift $\mathrm{d}\omega_P^{\mathcal{NE}} = S_P\mathrm{d}P$. The uncertainty in $\mathrm{d}\omega_P^{\mathcal{NE}}$ was calculated using standard error propagation methods using the uncertainties in the measured values of S_P and dP. All the shifts and uncertainties of this type are included in the final systematic error budget given in Table 1. If S_P was expected and observed to be consistent with zero, we did not apply a systematic correction associated with parameter P, but still computed an associated uncertainty. We include uncertainties of this type in the final systematic error budget in certain cases described below.

We identified several parameters that cause non-zero changes in $\omega^{\mathcal{NE}}$, which we discuss here. The first such contribution to systematic shifts arises from gradients of $\mathcal{B}_z$ along the z and y axes, $(\partial\mathcal{B}_z/\partial z)^{\mathrm{nr}}$ and $(\partial\mathcal{B}_z/\partial y)^{\mathrm{nr}}$ (Extended Data Fig. 2b). We understand the non-zero slopes associated with this set of parameters, $S_{\partial\mathcal{B}_z/\partial z} = \partial\omega^{\mathcal{NE}}/\partial(\partial\mathcal{B}_z/\partial z)$ and $S_{\partial\mathcal{B}_z/\partial y} = \partial\omega^{\mathcal{NE}}/\partial(\partial\mathcal{B}_z/\partial y)$, as follows. Because the spin precession phase, ϕ, is proportional to $\mathcal{B}_z$ (for $d_e = 0$; see equation (1)), a gradient $\partial\mathcal{B}_z/\partial z$ ($\partial\mathcal{B}_z/\partial y$), together with a translation of the centre of mass of the detected molecular beam along the direction of the gradient, dz_{cm} (dy_{cm}), can create a shift in the measured precession frequency, $\mathrm{d}\omega \propto (\partial\mathcal{B}_z/\partial z)\mathrm{d}z_{\mathrm{cm}}$ ($\mathrm{d}\omega \propto (\partial\mathcal{B}_z/\partial y)\mathrm{d}y_{\mathrm{cm}}$) (Extended Data Fig. 2a). We identified two separate effects that can cause such translations in our system: one that arises from a non-reversing electric field, $\mathcal{E}^{\mathrm{nr}}$, and one that arises from gradients in such a field, $\partial\mathcal{E}^{\mathrm{nr}}/\partial z$ and $\partial\mathcal{E}^{\mathrm{nr}}/\partial y$. Both effects are associated with incomplete laser excitation, and each can occur in both the STIRAP and probe laser beams. As described in detail in Supplementary Information, the systematic error requires a non-zero value of the readout-laser detuning (Δ), and a non-zero value of the STIRAP two-photon detuning (δ).

For the EDM dataset, we minimized the magnitudes of both slopes, $S_{\partial\mathcal{B}_z/\partial z}$ and $S_{\partial\mathcal{B}_z/\partial y}$, by tuning the readout laser so that $\Delta = 0$ and the STIRAP lasers so that $\delta = 0$ (Extended Data Fig. 2a). The residual slopes under these optimized conditions were then measured by applying large values of $\partial\mathcal{B}_z/\partial z$ and $\partial\mathcal{B}_z/\partial y$ and were found to be consistent with zero. To ensure that the imperfections leading to non-zero values of these slopes (namely, a combination of δ, Δ, $\mathcal{E}^{\mathrm{nr}}$, $\partial\mathcal{E}^{\mathrm{nr}}/\partial z$ or $\partial\mathcal{E}^{\mathrm{nr}}/\partial y$) did not drift to large values, these slopes were monitored at regular intervals throughout the EDM data collection (Extended Data Fig. 1e). Finally, the ambient values of $\partial\mathcal{B}_z/\partial z$ and $\partial\mathcal{B}_z/\partial y$ during the acquisition of the EDM dataset were minimized to less than $1\,\mu\mathrm{G\,cm^{-1}}$ using the magnetic-field coils (Fig. 2). These field gradients were monitored twice daily using in situ magnetometers near the molecular beam; additional offline measurements were made before

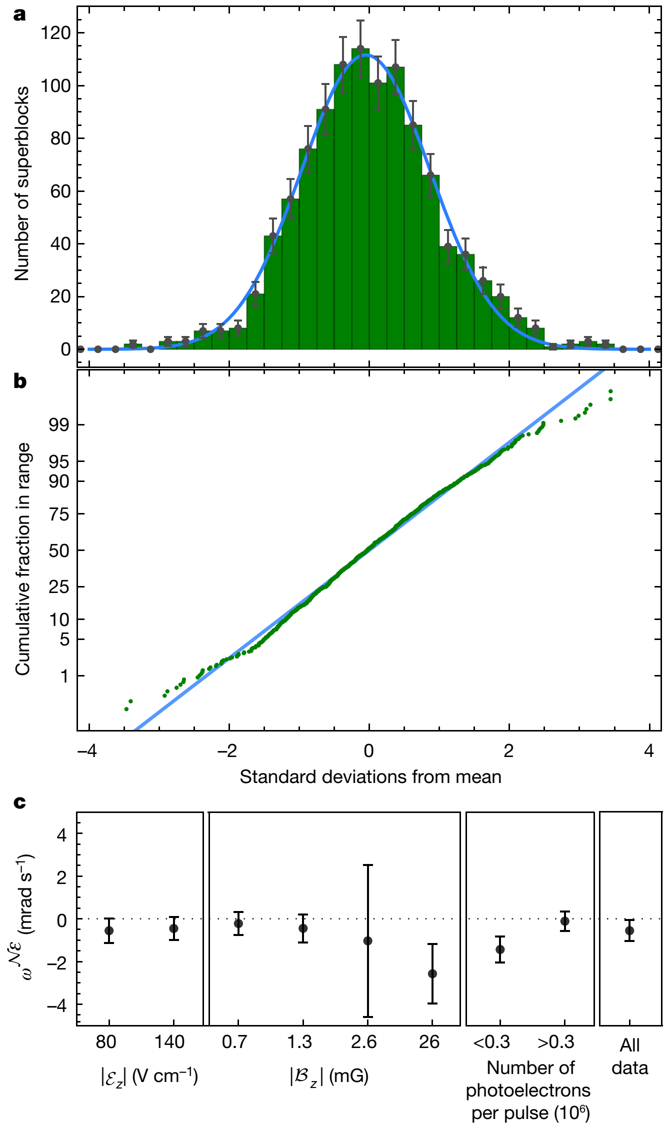


Fig. 3 | Statistics of the EDM dataset. a, Histogram of centred and normalized $\omega^{\mathcal{NE}}$ superblock values, that is, $(\omega^{\mathcal{NE}} - \langle\omega^{\mathcal{NE}}\rangle)/\sigma_{\omega^{\mathcal{NE}}}$. Here, $\langle\omega^{\mathcal{NE}}\rangle$ is the mean of $\omega^{\mathcal{NE}}$ over the dataset, $\sigma_{\omega^{\mathcal{NE}}} = \sigma^{\mathrm{sn}}_{\omega^{\mathcal{NE}}}\sqrt{\chi_{\mathrm{r}}^2(\mathcal{B})}$, where $\sigma^{\mathrm{sn}}_{\omega^{\mathcal{NE}}}$ is the superblock uncertainty propagated from 'groups' that is due to the shot noise, and $\chi_{\mathrm{r}}^2(\mathcal{B})$ is the reduced χ^2 value for the sets of superblocks for a given magnetic-field magnitude. Error bars indicate the standard deviation in the bin expected from a Poisson distribution. The blue line shows a Gaussian fit to the histogram. **b**, Normal probability plot (green points) compared with a normal distribution (blue line). Deviations from the line outside $\pm 1.5\sigma$ indicate more data points in the tails of the distribution than expected from a normal distribution. **c**, Values of $\omega^{\mathcal{NE}}$, grouped according to $|\mathcal{E}_z|$, $|\mathcal{B}_z|$, the block-averaged number of photoelectrons per pulse, and combined for all states. Error bars correspond to 1σ (68% confidence interval).

and after the acquisition of the EDM dataset by translating the position of the magnetometers along the molecular beam path. We include in the systematic error budget (Table 1) a contribution calculated from the values of the measured systematic slope $S_{\partial\mathcal{B}_z/\partial z}$ ($S_{\partial\mathcal{B}_z/\partial y}$) and the ambient $\partial\mathcal{B}_z/\partial z$ ($\partial\mathcal{B}_z/\partial y$).

The next parameter that contributes to systematic shifts is associated with an ellipticity gradient across the spatial profile of the STIRAP H–C laser beam. In practice, we control the size of this ellipticity gradient by using a half-waveplate to change the angle, $\theta_{\mathrm{ST}}^{\mathrm{H-C}}$, between the original polarization of the H–C laser and the average birefringence axis. As described in detail in Supplementary Information, an ambient

Table 1 | Systematic shifts for $\omega^{\mathcal{NE}}$ and their statistical uncertainties

Parameter	Shift	Uncertainty
$\partial\mathcal{B}_z/\partial z$ and $\partial\mathcal{B}_z/\partial y$	7	59
$\omega_{\mathrm{ST}}^{\mathcal{NE}}$ (via $\theta_{\mathrm{ST}}^{\mathrm{H-C}}$)	0	1
$P_{\mathrm{ref}}^{\mathcal{NE}}$	–	109
$\mathcal{E}^{\mathrm{nr}}$	−56	140
$\lvert\mathcal{C}\rvert^{\mathcal{NE}}$ and $\lvert\mathcal{C}\rvert^{\mathcal{NEB}}$	77	125
$\omega^{\mathcal{E}}$ (via $\mathcal{B}_z^{\mathcal{E}}$)	1	1
Other magnetic-field gradients (4)	–	134
Non-reversing magnetic field, $\mathcal{B}_z^{\mathrm{nr}}$	–	106
Transverse magnetic fields, $\mathcal{B}_x^{\mathrm{nr}}, \mathcal{B}_y^{\mathrm{nr}}$	–	92
Refinement- and readout-laser detunings	–	76
$\tilde{\mathcal{N}}$-correlated laser detuning, $\Delta^{\mathcal{N}}$	–	48
Total systematic	29	310
Statistical uncertainty		373
Total uncertainty		486

Values are shown in μrad s^{-1}. All uncertainties are added in quadrature. For $\mathcal{E}_{\mathrm{eff}} = 78$ GV cm^{-1}, $d_e = 10^{-30}e$ cm corresponds to $\lvert\omega^{\mathcal{NE}}\rvert = \mathcal{E}_{\mathrm{eff}} d_e/\hbar = 119$ μrad s^{-1}.

birefringence gradient, in combination with a finite value of the refinement-laser beam attenuation, A_{ref}, and a non-zero $\mathcal{E}^{\mathrm{nr}}$ leads to a non-zero value of $S_{\theta_{\mathrm{ST}}^{\mathrm{H-C}}} = \partial\omega^{\mathcal{NE}}/\partial\theta_{\mathrm{ST}}^{\mathrm{H-C}} = (\partial\omega_{\mathrm{ST}}^{\mathcal{NE}}/\partial\theta_{\mathrm{ST}}^{\mathrm{H-C}})/A_{\mathrm{ref}}$.

Throughout the acquisition of the EDM dataset, we measured the slope $S_{\theta_{\mathrm{ST}}^{\mathrm{H-C}}}$ by applying a large $\theta_{\mathrm{ST}}^{\mathrm{H-C}}$ and measuring the value of $\omega^{\mathcal{NE}}$ that survives refinement. This value is consistent with zero, directly bounding the attenuation under ordinary conditions to $A_{\mathrm{ref}} > 10^4$. We measured the value of $\theta_{\mathrm{ST}}^{\mathrm{H-C}}$ with the following procedure. By tuning the power of the refinement laser, P_{ref}, to zero so that $A_{\mathrm{ref}} = 1$, we observed a contribution to the precession frequency associated with the STIRAP state-preparation laser beams, ω_{ST}. Consistent with the ellipticity-gradient model described above, under these conditions we also observed an $\mathcal{NE}$-correlated component, $\omega_{\mathrm{ST}}^{\mathcal{NE}}$, resulting from the combination of a.c. Stark-shift effects and a non-zero $\delta^{\mathcal{NE}}$ (caused by the residual ambient $\mathcal{E}^{\mathrm{nr}}$). The slope $\partial\omega_{\mathrm{ST}}^{\mathcal{NE}}/\partial\theta_{\mathrm{ST}}^{\mathrm{H-C}}$ was calibrated by setting $P_{\mathrm{ref}} = 0$ and measuring the dependence of $\omega_{\mathrm{ST}}^{\mathcal{NE}}$ on an exaggerated $\theta_{\mathrm{ST}}^{\mathrm{H-C}}$. Finally, the value of $\theta_{\mathrm{ST}}^{\mathrm{H-C}}$ was found from the relation $\theta_{\mathrm{ST}}^{\mathrm{H-C}} = \omega_{\mathrm{ST}}^{\mathcal{NE}}/(\partial\omega_{\mathrm{ST}}^{\mathcal{NE}}/\partial\theta_{\mathrm{ST}}^{\mathrm{H-C}})$. To minimize the ellipticity gradient, we set $\theta_{\mathrm{ST}}^{\mathrm{H-C}}$ to the value that was found to minimize $\omega_{\mathrm{ST}}^{\mathcal{NE}}$. Both $\omega_{\mathrm{ST}}^{\mathcal{NE}}$ and the slope $S_{\theta_{\mathrm{ST}}^{\mathrm{H-C}}}$ were monitored at regular intervals throughout the acquisition of the EDM dataset (Extended Data Fig. 1e). The measured values of the systematic slope $S_{\theta_{\mathrm{ST}}^{\mathrm{H-C}}}$ and the residual $\theta_{\mathrm{ST}}^{\mathrm{H-C}}$ were used to compute the contribution of the STIRAP lasers to the systematic error budget (Table 1).

Another parameter that contributes to a systematic shift of $\omega^{\mathcal{NE}}$ is an $\tilde{\mathcal{N}}\tilde{\mathcal{E}}$-correlated component of the power of the refinement beam, defined by $P_{\mathrm{ref}} = P_{\mathrm{ref}}^{\mathrm{nr}} + \tilde{\mathcal{N}}\tilde{\mathcal{E}}P_{\mathrm{ref}}^{\mathcal{NE}}$. As illustrated in Supplementary Information, a misalignment between the $\boldsymbol{\epsilon}_{\mathrm{ref}}$ and $\boldsymbol{S}_{\mathrm{ST}}$ polarization vectors, $\theta_{\mathrm{ST}}^{\mathrm{ref}}$, leads to a non-zero value in the slope $S_{P_{\mathrm{ref}}^{\mathcal{NE}}} = \partial\omega^{\mathcal{NE}}/\partial P_{\mathrm{ref}}^{\mathcal{NE}}$.

For the EDM dataset, we minimized the magnitude of $S_{P_{\mathrm{ref}}^{\mathcal{NE}}}$ by tuning $\theta_{\mathrm{ST}}^{\mathrm{ref}}$ to zero via a half-waveplate in the refinement-laser beam. We did not observe clear evidence of a non-zero $P_{\mathrm{ref}}^{\mathcal{NE}}$ component in our EDM dataset. However, we put a limit on its possible size throughout the acquisition of the EDM dataset by placing bounds on the offset of $\omega^{\mathcal{NEB}}$, which has a strong linear dependence on $P_{\mathrm{ref}}^{\mathcal{NE}}$ owing to a.c. Stark-shift effects[1,9]. The $\partial\omega^{\mathcal{NE}}/\partial P_{\mathrm{ref}}^{\mathcal{NE}}$ slope was monitored regularly throughout the acquisition of the EDM dataset (Extended Data Fig. 1e). We used the measured upper limit of $P_{\mathrm{ref}}^{\mathcal{NE}}$ and the value of $\partial\omega^{\mathcal{NE}}/\partial P_{\mathrm{ref}}^{\mathcal{NE}}$ to calculate a contribution of $P_{\mathrm{ref}}^{\mathcal{NE}}$ to the systematic error budget (Table 1).

The next parameter that contributes to the systematic error budget is $\mathcal{E}^{\mathrm{nr}}$, which has already been discussed as one of the parameters needed to induce the $\partial\mathcal{B}_z/\partial z$ ($\partial\mathcal{B}_z/\partial y$) and $\omega_{\mathrm{ST}}^{\mathcal{NE}}$ systematic effects. However, an additional contribution arises from imperfections in the ellipticity gradients of the refinement and readout lasers in combination with $\mathcal{E}^{\mathrm{nr}}$, which was one of the dominant systematic effects in ACME I[1,9]. By applying a large value of $\mathcal{E}^{\mathrm{nr}}$, we measured $S_{\mathcal{E}^{\mathrm{nr}}} = \partial\omega^{\mathcal{NE}}/\partial\mathcal{E}^{\mathrm{nr}}$ regularly throughout the acquisition of the EDM dataset (Extended Data Fig. 1e). $\mathcal{E}^{\mathrm{nr}}$ and its gradients in the precession region, $\partial\mathcal{E}^{\mathrm{nr}}/\partial z$ and $\partial\mathcal{E}^{\mathrm{nr}}/\partial y$, were measured every two weeks during the acquisition of the EDM dataset using a mapping method based on microwave spectroscopy[9]. We include in the systematic error budget (Table 1) a contribution of this $\mathcal{E}^{\mathrm{nr}}$ systematic effect based on $S_{\mathcal{E}^{\mathrm{nr}}}$ and the measured ambient $\mathcal{E}^{\mathrm{nr}}$.

The next contribution to the systematic error arises from imperfections in the spin-measurement contrast, $\mathcal{C}$. As described in detail in Supplementary Information, we observed correlations $S_{\lvert\mathcal{C}\rvert^u} = \partial\omega^{\mathcal{NE}}/\partial\lvert\mathcal{C}\rvert^u$ with two contrast channels, $\lvert\mathcal{C}\rvert^{\mathcal{NE}}$ and $\lvert\mathcal{C}\rvert^{\mathcal{NEB}}$. Although the average values $\langle\lvert\mathcal{C}\rvert^{\mathcal{NE}}\rangle$ and $\langle\lvert\mathcal{C}\rvert^{\mathcal{NEB}}\rangle$ of the corresponding contrast channels are consistent with zero in the EDM dataset, we include in our error budget a limit on their possible contributions extracted from $S_{\lvert\mathcal{C}\rvert^{\mathcal{NE}}}$ ($S_{\lvert\mathcal{C}\rvert^{\mathcal{NEB}}}$) and $\langle\lvert\mathcal{C}\rvert^{\mathcal{NE}}\rangle$ ($\langle\lvert\mathcal{C}\rvert^{\mathcal{NEB}}\rangle$) (Table 1).

The last parameter observed to generate a systematic shift was $\omega^{\mathcal{E}}$, which can result from leakage-current, motional-magnetic-field ($\boldsymbol{v}\times\boldsymbol{\mathcal{E}}$) and geometric-phase effects[19]. To measure the slope $S_{\omega^{\mathcal{E}}} = \partial\omega^{\mathcal{NE}}/\partial\omega^{\mathcal{E}}$, we apply an $\tilde{\mathcal{E}}$-correlated component of the magnetic field, $\mathcal{B}_z^{\mathcal{E}}$, which creates a large artificial $\omega^{\mathcal{E}}$. $S_{\omega^{\mathcal{E}}}$ is a measure of the suppression of any residual value of $\omega^{\mathcal{E}}$ by the $\tilde{\mathcal{N}}$ switch[20,21]. The mean value of $\omega^{\mathcal{E}}$ in the EDM dataset, $\langle\omega^{\mathcal{E}}\rangle$, was measured to be consistent with zero. We place a limit on the possible contributions from $\omega^{\mathcal{E}}$ effects using the measured values of $S_{\omega^{\mathcal{E}}}$ and $\langle\omega^{\mathcal{E}}\rangle$ (Table 1).

In addition to the above effects, we include in our systematic error budget possible contributions from the following parameters (all closely related to the parameters observed to cause a non-zero $\omega^{\mathcal{NE}}$ shift in our measurement): residual (non-reversing) magnetic fields (along all three directions), all additional first-order magnetic-field gradients ($\partial\mathcal{B}_x/\partial x, \partial\mathcal{B}_y/\partial y, \partial\mathcal{B}_y/\partial x, \partial\mathcal{B}_z/\partial x$), refinement- and readout-laser detunings and the differential detuning between the two experimental $\tilde{\mathcal{N}}$ states, $\Delta^{\mathcal{N}}$.

Results and conclusions

The result of this second-generation EDM measurement using ThO is $\omega^{\mathcal{NE}} = -510 \pm 373_{\mathrm{stat}} \pm 310_{\mathrm{syst}}$ μrad s^{-1}. Using $d_e = -\hbar\omega^{\mathcal{NE}}/\mathcal{E}_{\mathrm{eff}}$ and[16,17] $\mathcal{E}_{\mathrm{eff}} \approx 78$ GV cm^{-1} results in

$$d_e = (4.3 \pm 3.1_{\mathrm{stat}} \pm 2.6_{\mathrm{syst}}) \times 10^{-30} e\ \mathrm{cm} \tag{4}$$

where the combined statistical and systematic uncertainty, $\sigma_{d_e} = 4.0 \times 10^{-30} e$ cm, is a factor of 12 smaller than the previous best result, from ACME I[1,9].

An upper limit on $\lvert d_e\rvert$ is computed by applying the Feldman–Cousins prescription[9,33] to a folded normal distribution, which yields

$$\lvert d_e\rvert < 1.1 \times 10^{-29} e\ \mathrm{cm} \tag{5}$$

at 90% confidence level. This is 8.6 times smaller than the best previous limit, from ACME I[1,9]. Because paramagnetic molecules are sensitive to multiple time-reversal-symmetry-violating effects[34], our measurement can be more generally interpreted as $\hbar\omega^{\mathcal{NE}} = -d_e\mathcal{E}_{\mathrm{eff}} + W_{\mathrm{S}}C_{\mathrm{S}}$, where C_{S} is a dimensionless time-reversal-symmetry-violating electron–nucleon coupling parameter and $W_{\mathrm{S}} = -2\pi\hbar \times 282$ kHz is a molecule-specific constant[16,17,35]. For the d_e limit given above, we assume $C_{\mathrm{S}} = 0$. Assuming $d_e = 0$ instead gives $\lvert C_{\mathrm{S}}\rvert < 7.3 \times 10^{-10}$ (90% confidence level).

Because the values of d_e and C_{S} predicted by the standard model are many orders of magnitude below our sensitivity[2,3], this measurement is a background-free probe for new physics beyond the standard model. Nearly every extension of the standard model[4–6] introduces the possibility for new particles and new time-reversal-symmetry-violating phases, ϕ_{T}, that can lead to measurable EDMs. Within typical extensions of the standard model, an EDM arising from new particles

with rest-mass energy Λ in an n-loop Feynman diagram will have a size of[8,14,36]

$$\frac{d_e}{e} \approx \kappa \left(\frac{\alpha_{\mathrm{eff}}}{2\pi}\right)^n \left(\frac{m_e c^2}{\Lambda^2}\right) \sin(\phi_{\mathrm{T}})(\hbar c) \qquad (6)$$

where α_{eff} ($\alpha_{\mathrm{eff}} = 4/137$ for electroweak interactions) encodes the strength with which the electron couples to new particles, m_e is the electron mass and $\kappa \approx 0.1$–1 is a dimensionless prefactor whose value depends on the specific model of new physics. In typical models, where d_e is produced by one- or two-loop diagrams, for $\sin(\phi_{\mathrm{T}}) \approx 1$ our result typically limits time-reversal-symmetry-violating new physics to energy scales above $\Lambda \approx 30$ TeV or $\Lambda \approx 3$ TeV, respectively[4–8].

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0599-8.

Received: 12 June 2018; Accepted: 20 August 2018;
Published online 17 October 2018.

1. Baron, J. et al. Order of magnitude smaller limit on the electric dipole moment of the electron. *Science* **343**, 269–272 (2014).
2. Pospelov, M. E. & Khriplovich, I. B. Electric dipole moment of the W boson and the electron in the Kobayashi–Maskawa model. *Sov. J. Nucl. Phys*. **53**, 638–640 (1991).
3. Pospelov, M. & Ritz, A. CKM benchmarks for electron electric dipole moment experiments. *Phys. Rev. D* **89**, 056006 (2014).
4. Nakai, Y. & Reece, M. Electric dipole moments in natural super symmetry. *J. High Energy Phys*. **8**, 31 (2017).
5. Barr, S. M. A review of CP violation in atoms. *Int. J. Mod. Phys. A* **08**, 209–236 (1993).
6. Pospelov, M. & Ritz, A. Electric dipole moments as probes of new physics. *Ann. Phys*. **318**, 119–169 (2005).
7. Engel, J., Ramsey-Musolf, M. J. & van Kolck, U. Electric dipole moments of nucleons, nuclei, and atoms: the standard model and beyond. *Prog. Part. Nucl. Phys*. **71**, 21–74 (2013).
8. Bernreuther, W. & Suzuki, M. The electric dipole moment of the electron. *Rev. Mod. Phys*. **63**, 313–340 (1991).
9. ACME Collaboration et al. Methods, analysis, and the treatment of systematic errors for the electron electric dipole moment search in thorium monoxide. *New J. Phys*. **19**, 073029 (2016).
10. Hudson, J. J. et al. Improved measurement of the shape of the electron. *Nature* **473**, 493–496 (2011).
11. Kara, D. M. et al. Measurement of the electron's electric dipole moment using YbF molecules: methods and data analysis. *New J. Phys*. **14**, 103051 (2012).
12. Cairncross, W. B. et al. Precision measurement of the electron's electric dipole moment using trapped molecular ions. *Phys. Rev. Lett*. **119**, 153001 (2017).
13. Sandars, P. G. H. The electric dipole moment of an atom. *Phys. Lett*. **14**, 194–196 (1965).
14. Khriplovich, I. B. & Lamoreaux, S. K. *CP Violation Without Strangeness* (Springer, NewYork, 1997).
15. Commins, E. D. & DeMille, D. in *Lepton Dipole Moments* (eds Roberts, B. L. & Marciano, W. J.) Ch. 14 (World Scientific, Singapore, 2010).
16. Denis, M. & Fleig, T. In search of discrete symmetry violations beyond the standard model: thorium monoxide reloaded. *J. Chem. Phys*. **145**, 214307 (2016).
17. Skripnikov, L. V. Combined 4-component and relativistic pseudo potential study of ThO for the electron electric dipole moment search. *J. Chem. Phys*. **145**, 214301 (2016).
18. Vutha, A. C. et al. Search for the electric dipole moment of the electron with thorium monoxide. *J. Phys. B* **43**, 074007 (2010).
19. Regan, B. C., Commins, E. D., Schmidt, C. J. & DeMille, D. New limit on the electron electric dipole moment. *Phys. Rev. Lett*. **88**, 071805 (2002).
20. Bickman, S., Hamilton, P., Jiang, Y. & DeMille, D. Preparation and detection of states with simultaneous spin alignment and selectable molecular orientation in PbO. *Phys. Rev. A* **80**, 023418 (2009).
21. Eckel, S., Hamilton, P., Kirilov, E., Smith, H. W. & DeMille, D. Search for the electron electric dipole moment using -doublet levels in PbO. *Phys. Rev. A* **87**, 052130 (2013).
22. Kirilov, E. et al. Shot-noise-limited spin measurements in a pulsed molecular beam. *Phys. Rev. A* **88**, 013844 (2013).
23. Hutzler, N. R., Lu, H. I. & Doyle, J. M. The buffer gas beam: an intense, cold, and slow source for atoms and molecules. *Chem. Rev*. **112**, 4803–4827 (2012).
24. Hutzler, N. R. et al. A cryogenic beam of refractory, chemically reactive molecules with expansion cooling. *Phys. Chem. Chem. Phys*. **13**, 18976 (2011).
25. Patterson, D. & Doyle, J. M. Bright, guided molecular beam with hydrodynamic enhancement. *J. Chem. Phys*. **126**, 154307 (2007).
26. Panda, C. D. et al. Stimulated Raman adiabatic passage preparation of a coherent superposition of ThO $H^3\Delta_1$ states for an improved electron electric-dipole-moment measurement. *Phys. Rev. A* **93**, 052110 (2016).
27. Gray, H. R., Whitley, R. M. & Stroud, C. R. Coherent trapping of atomic populations. *Opt. Lett*. **3**, 218–220 (1978).
28. Kokkin, D. L., Steimle, T. C. & DeMille, D. Branching ratios and radiative lifetimes of the U, L, and i states of thorium oxide. *Phys. Rev. A* **90**, 062503 (2014).
29. Kokkin, D. L., Steimle, T. C. & DeMille, D. Characterization of the $I(|\Omega| = 1) - X^1\Sigma^+$ (0, 0) band of thorium oxide. *Phys. Rev. A* **91**, 042508 (2015).
30. Huber, P. J. Robust estimation of a location parameter. *Ann. Math. Stat*. **35**, 73–101 (1964).
31. Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Stat*. **7**, 1–26 (1979).
32. Efron, B. & Tibshirani, R. Bootstrap Methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci*. **1**, 54–75 (1986).
33. Feldman, G. J. & Cousins, R. D. Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D* **57**, 3873–3889 (1998).
34. Kozlov, M. G. & Labzowsky, L. N. Parity violation effects in diatomic molecules. *J. Phys. B* **28**, 1933–1961 (1995).
35. Dzuba, V. A., Flambaum, V. V. & Harabati, C. Relations between matrix elements of different weak interactions and interpretation of the parity-nonconserving and electron electric-dipole-moment measurements in atoms and molecules. *Phys. Rev. A* **84**, 052108 (2011).
36. Fortson, N., Sandars, P. & Barr, S. The search for a permanent electric dipole moment. *Phys. Today* **56**, 33–39 (2003).

Acknowledgements This work was supported by the NSF. J.H. was supported by the Department of Defense. D.G.A. was partially supported by the Amherst College Kellogg University Fellowship. We thank M. Reece and M. Schwartz for discussions and S. Cotreau, J. MacArthur and S. Sansone for technical support.

Reviewer information *Nature* thanks E. Hinds and Y. Shagam for their contribution to the peer review of this work.

Author contributions All authors contributed to one or more of the following areas: proposing, leading and running the experiment; design, construction, optimization and testing of the experimental apparatus and data acquisition system; setup and maintenance during the data runs; data analysis and extraction of physics results from measured traces; modelling and simulation of systematic errors; and the writing of this article. The corresponding authors are D.D., J.M.D. and G.G. (acme@physics.harvard.edu).

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0599-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0599-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

ACME Collaboration
V. Andreev[1,5], D. G. Ang[1], D. DeMille[2]*, J. M. Doyle[1]*, G. Gabrielse[1,3]*, J. Haefner[1], N. R. Hutzler[1,4], Z. Lasner[2], C. Meisenhelder[1], B. R. O'Leary[2], C. D. Panda[1], A. D. West[2,6], E. P. West[1,6] & X. Wu[1,2]

[1]Department of Physics, Harvard University, Cambridge, MA, USA. [2]Department of Physics, Yale University, New Haven, CT, USA. [3]Center for Fundamental Physics, Northwestern University, Evanston, IL, USA. [4]Division of Physics, Mathematics, and Astronomy, California Institute of Technology, Pasadena, CA, USA. [5]Present address: Max Planck Institute of Quantum Optics, Garching, Germany. [6]Present address: Department of Physics and Astronomy, UCLA, Los Angeles, CA, USA. *e-mail: acme@physics.harvard.edu

METHODS

Apparatus. To describe the experiment (and its imperfections) in detail, we define a coordinate system in which $\hat{z}$ is the direction of the applied electric field (pointing from east to west in our laboratory), $\hat{y}$ (which is along $\hat{z} \times \langle v \rangle$, where $\langle v \rangle$ is the average molecular velocity) points approximately downwards, and $\hat{x} = \hat{y} \times \hat{z}$ is approximately parallel to $\langle v \rangle$ (Fig. 2). This system is used throughout the main body of the article.

After leaving the beam source, the molecules have a thermal distribution of rotational states at 4 K and reside mostly (>70%) in the $J = 0$–2 rotational levels. We use two stages of optical pumping for 'rotational cooling', namely, to enhance the population in the ground rotational level $|X, J = 0, P = +1\rangle$. The first stage is performed in an electric field of about 0 V cm^{-1}, using 5–7 passes of a laser beam resonant with the $|X, J = 2, P = +1\rangle \leftrightarrow |C, J = 1, P = -1\rangle$ transition. Each pass has orthogonal polarization to the previous one, addressing all possible M states in $|X, J = 2, P = +1\rangle$. Owing to the parity and angular momentum selection rules, this results in optical pumping of population from $|X, J = 2, P = +1\rangle$ to $|X, J = 0, P = +1\rangle$. The second stage is performed in an applied electric field of about 100 V cm^{-1}, which mixes the opposite-parity excited states $|C, J = 1, P = \pm 1, M = \pm 1\rangle$. A multipass, alternating polarization laser beam drives the $|X, J = 1, P = -1\rangle \leftrightarrow |C, J = 1, P = \text{mixed}, M = \pm 1\rangle$ transition, partially transferring population from $|X, J = 1\rangle$ to $|X, J = 0\rangle$. These two combined rotational-cooling steps increase the population in the $|X, J = 0\rangle$ state by a factor of 2.5.

The molecules then pass through fixed collimating apertures before entering the magnetically shielded spin precession region, where the $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\mathcal{E}}$ fields are applied. The electric field is produced by a pair of parallel fused silica plates coated with a thin layer (20 nm) of indium tin oxide (ITO) on one side and anti-reflection coating on the other side[37]. The ITO-coated sides face each other with a gap of 45 mm between them and are connected to low-noise voltage supplies.

A spatial map of the electric-field magnitude was measured by performing microwave spectroscopy on the ThO molecules[9]. This measurement indicated that the non-reversing component of the electric field had a varying magnitude of $|\mathcal{E}^{\rm nr}| \approx 1$–5 mV cm^{-1} at different points along the propagation direction of the molecular beam, x. We measured the spatial dependence of $\mathcal{E}^{\rm nr}$ in z and y by selectively blocking half of the STIRAP state-preparation and readout laser beams, respectively.

The vacuum chamber that houses the spin precession region is surrounded by five layers of mu-metal shielding. The coil design is optimized to create a uniform magnetic field along z. Additional coils allow us to apply magnetic-field offsets in the transverse directions (x and y), as well as all first-order gradients ($\partial\mathcal{B}_z/\partial z$, $\partial\mathcal{B}_z/\partial y$, $\partial\mathcal{B}_x/\partial x$, $\partial\mathcal{B}_y/\partial y$, $\partial\mathcal{B}_y/\partial x$, $\partial\mathcal{B}_z/\partial x$), for systematic error checks. The magnetic field is monitored by four three-axis fluxgate magnetometers, which are placed inside pockets inset in the vacuum chamber, 20–30 cm from the molecular beam. The electronic offset that is inherent to the fluxgate magnetometers is subtracted by rotating them in situ by 180°, and the position of each magnetometer can be translated along one axis. These magnetometers are used to infer changes in the magnetic field as well as information about its gradients. The magnetic field is also mapped before and after the acquisition of the EDM dataset by sliding a three-axis fluxgate magnetometer down the beamline (along $\hat{x}$) at the position of the molecules and at positions offset vertically (along $\hat{y}$). Measurement of the gradients along $\hat{x}$ and $\hat{y}$, along with Maxwell's equations, allow us to also infer the gradients along $\hat{z}$, where the mechanical geometry of the field plates prevent a direct measurement.

The STIRAP lasers travel vertically through the experimental setup, between the field plates. They are launched from the beamshaping optics at the top of the setup, which overlaps and focuses the two laser beams (waists of about 150 μm) at the position of the molecular beam[26]. The refinement and readout beams travel horizontally through the field plates, so all stages of the spin precession measurement are performed in a uniform electric field.

The STIRAP light originates from external cavity diode lasers (ECDLs) whose frequencies are locked to the resonance of an ultralow expansion glass (ULE) cavity. A linear drift of 7 kHz d^{-1}, due to the mechanical relaxation of the ULE spacer, is measured using a stabilized frequency comb and is corrected for using acousto-optic modulators (AOMs)[26]. The refinement and readout lasers both derive from the same Ti:sapphire (Ti:S) laser (703 nm). We switch the Ti:S laser frequency between the two $\tilde{\mathcal{N}}$ states by tuning the length of the laser cavity using piezoelectric elements. The Ti:S laser is locked to the ULE cavity via a transfer lock to a separate 703 nm ECDL. We address the two $\tilde{\mathcal{P}}$ states by shifting the frequency of the readout laser with the AOMs. Unlike in ACME I[1,9], a global rotation of both the refinement- and readout-laser polarizations, $\tilde{\mathcal{G}}$, cannot be implemented because the spin alignment is already fixed by the polarization of the Stokes STIRAP laser to be nominally along $\hat{x}$[26].

To normalize against the changing molecule number, we alternate the readout-laser polarization fast enough so that each molecule is reliably projected onto one of the two spin-alignment directions, with a probability determined by the orientation of its spin during its time of flight through the laser beam[22]. To do so, we overlap two laser beams with orthogonal $\hat{X}$ and $\hat{Y}$ polarizations, which we switch on and off rapidly using the AOMs. The $\hat{X}$ and $\hat{Y}$ pulses each have a duration of 1.9 μs, with a 0.6 μs delay between them to minimize the overlap of signal due to the finite lifetime of the I state (115 ns)[29] between successive pulses (Extended Data Fig. 1a).

Fluorescence photons travel through the transparent field plates and are focused by one of eight lenses (four behind each field plate) into one of eight bent fused silica lightguides. Each lightguide carries the fluorescence photons to one of eight PMTs outside the magnetic shielding. The PMTs are optimized to detect fluorescence at 512 nm (approximately 25% quantum efficiency).

The PMT photocurrents are amplified and then recorded by a 14-bit digitizer operating at 16 million samples per second. All fast-timing (>10 Hz) electronics are controlled by a timing generator with jitter that is less than one sampling period of the digitizer. The digitizer signal is recorded by a computer, which communicates with a second computer that controls the slow switches (<10 Hz).

Statistics. The total run time for the collection of the EDM dataset was about 500 h; of these, about 350 h produced data that were used to compute the EDM and about 150 h were used for interleaved systematic-error checks (Extended Data Fig. 1). We also paused the experiment for about 8 h every 24 h (typically during the night) to thermally cycle the beam source to remove neon ice buildup.

Our robust M-estimator analysis[30,32] was performed using different weighting functions, such as the Huber, Hampel and Tukey functions. To obtain the quoted results we used the Huber weighting function because of its simplicity and wide use; other choices would change the mean and its uncertainty by only a few per cent. This procedure also yields results consistent with those found using alternative methods, such as directly scaling the error bars by χ_r^2 or forming the 5% trimmed mean of bootstrapped data[38]. Although our d_e limit is computed using the Feldman–Cousins prescription[9,33], previous EDM experiments[10,12,19] have reported limits based on a direct folded Gaussian distribution. To facilitate comparison with those experiments, we note that our limit computed in this way would be $|d_e| < 9.6 \times 10^{-30} e$ cm (90% confidence level).

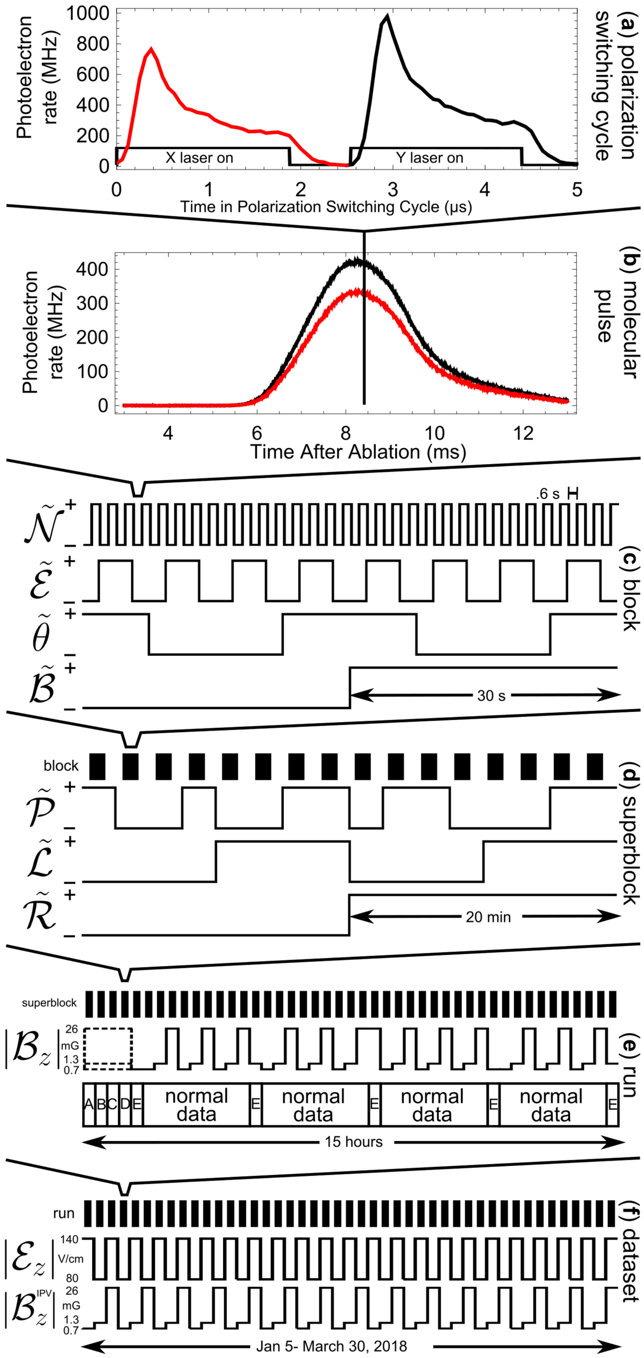
Future improvements. We believe that substantial improvements in sensitivity are possible with further development of the ACME technique. Detecting multiple photons from each molecule via optical cycling[39] could increase the experiment detection efficiency by an order of magnitude. Electric or magnetic focusing of the ThO molecular beam could increase the number of measured molecules by another order of magnitude, whereas improvements in cryogenic buffer-gas beam-source technology could give further gains. We are exploring various methods, such as faster switching between the addressed $\tilde{\mathcal{N}}$ states, to reduce the excess noise observed in ACME II. The dominant systematic errors that we observed can be further suppressed by improved magnetic-field control and reduced polarization gradients in the laser beams.

Code availability. The computer codes used for the analysis of the data are available from the corresponding authors on reasonable request.

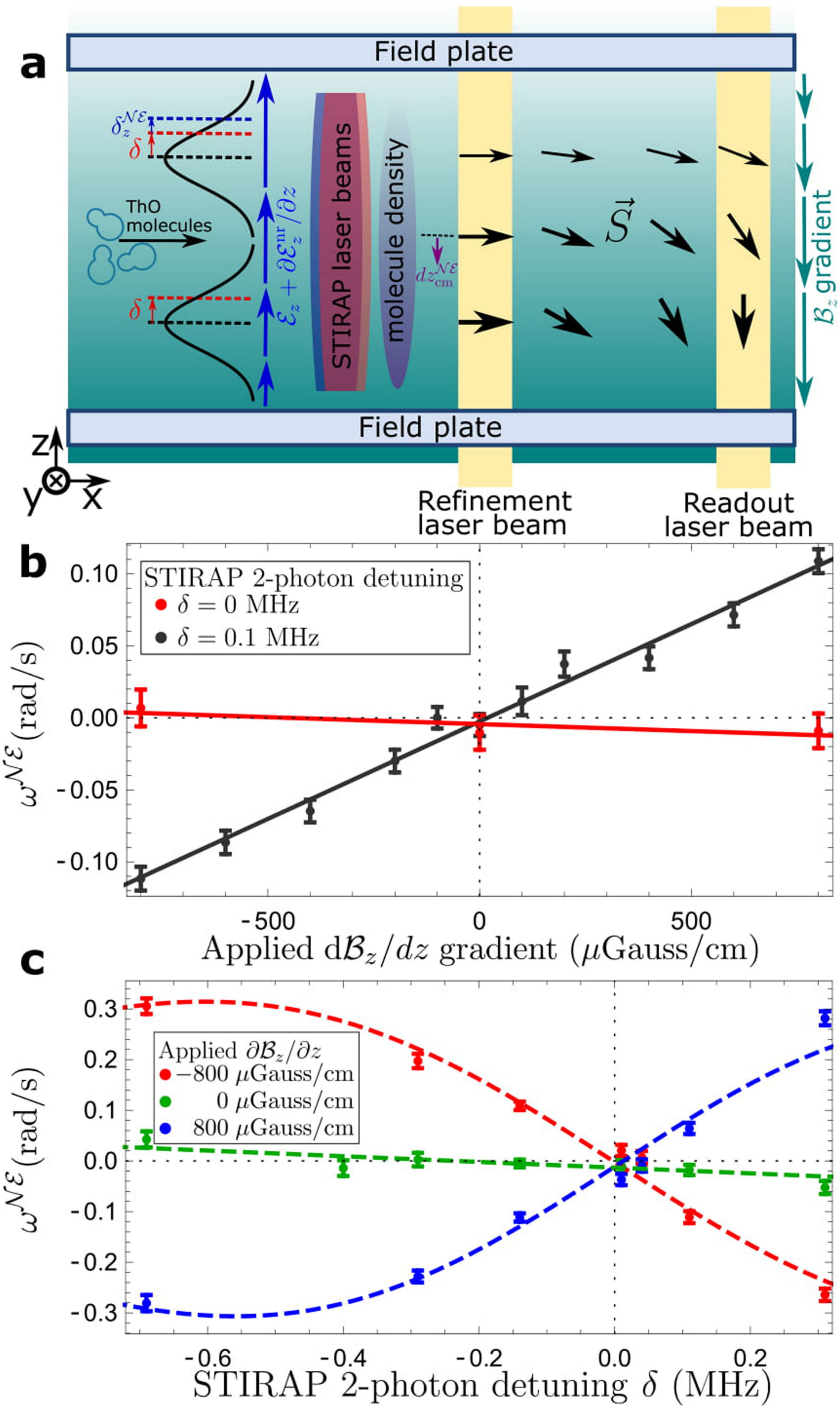
Data availability

The data that support the conclusions of this article are available from the corresponding authors on reasonable request.

37. Andreev, V., Panda, C. D., Hess, P. W., Spaun, B. & Gabrielse, G. A self-calibrating polarimeter to measure Stokes parameters. Preprint at https://arxiv.org/abs/1703.00963 (2017).
38. Kenney, J. F. & Keeping, E. S. *Mathematics of Statistics: Part One* 4th edn (Chapman & Hall, London, 1954).
39. Shuman, E. S., Barry, J. F., Glenn, D. R. & DeMille, D. Radiative force from optical cycling on a diatomic molecule. *Phys. Rev. Lett.* **103**, 223001 (2009).



Extended Data Fig. 1 | Switching timescales. a, Fluorescence signal amplitude versus time in an $\hat{X}$, $\hat{Y}$ polarization cycle. The red line corresponds to the signal from the $\hat{X}$-polarization laser and the black line to the signal from the $\hat{Y}$-polarization laser. **b**, Measured molecular trace (25 averaged pulses) versus time. Signal averaged over the entire $\hat{X}$, $\hat{Y}$ polarization cycles shown in **a** are shown in red and black for the $\hat{X}$ and $\hat{Y}$ laser polarizations, respectively. **c**, Switches performed within a block. The $\tilde{\mathcal{N}}$ and $\tilde{\mathcal{B}}$ switches randomly alternate between a $(-+)$ and a $(+-)$ pattern, and the $\tilde{\mathcal{E}}$ and $\tilde{\theta}$ switches randomly alternate between $(-++-)$ and $(+--+)$ between blocks. **d**, Switches performed within a superblock. The $\tilde{\mathcal{P}}$-state order is selected randomly, while $\tilde{\mathcal{L}}$ and $\tilde{\mathcal{R}}$ are deterministic. **e**, Run-data structure. We alternate between 'normal' EDM data, taken at three values of $|\mathcal{B}_z|$, and monitoring of known systematic effects by performing intentional parameter variations (IPVs). For several days data were taken with $|\mathcal{B}_z| = 2.6$ mG instead of $|\mathcal{B}_z| = 0.7$ mG, which is shown in the figure. Each IPV corresponds to one superblock, where a control parameter (A–E) is deliberately offset from its ideal value. Here, $A = P_{\mathrm{ref}}$ (the refinement beam is completely blocked, to determine the intrinsic $\omega_{\mathrm{ST}}^{\mathcal{N}\mathcal{E}}$), $B = \mathcal{E}^{\mathrm{nr}}$, $C = P^{\mathcal{N}\mathcal{E}}$, $D = \phi_{\mathrm{ST}}^{\mathcal{N}\mathcal{E}}$ and $E = \partial\mathcal{B}_z/\partial z$. The magnetic-field magnitude for the IPV of parameter E was varied between three experimental values within a run. **f**, The EDM dataset. The electric-field magnitude was varied from day to day. The magnetic-field magnitude for the IPVs for parameters A, B, C and D was varied between three experimental values.



Extended Data Fig. 2 | The $\partial\mathcal{B}_z/\partial z \times \delta \times \partial\mathcal{E}^{\mathrm{nr}}/\partial z$ systematic error. **a**, A $\partial\mathcal{E}^{\mathrm{nr}}/\partial z$ gradient (blue arrows) causes a z-dependent two-photon detuning correlated with $\mathcal{NE}$ ($\delta_z^{\mathcal{NE}}$), due to the Stark shift $D\mathcal{E}$. When $\delta \neq 0$, the combination of a non-zero $\delta_z^{\mathcal{NE}}$ and a dependence of the STIRAP efficiency on the two-photon detuning, $\partial\eta/\partial\delta$ (shown as black lines), acts to translate the detected molecular cloud (purple gradient ellipse) position by $\mathrm{d}z_{\mathrm{cm}}^{\mathcal{NE}}$ (purple arrow). A non-zero $\partial\mathcal{B}_z/\partial z$ (teal-colour gradient) causes molecules to accumulate more (less) precession phase if their position has a smaller (larger) z coordinate. The effects combine to create the dependence of $\omega^{\mathcal{NE}}$ on $\partial\mathcal{B}_z/\partial z$. The scales are exaggerated for clarity. **b**, The effect of changing the STIRAP two-photon detuning, δ, on the $\omega^{\mathcal{NE}}$ versus $\partial\mathcal{B}_z/\partial z$. We note that the slope $\partial\omega^{\mathcal{NE}}/\partial(\partial\mathcal{B}_z/\partial z)$ is consistent with zero when δ is set to zero. **c**, Dependence of $\omega^{\mathcal{NE}}$ on δ and $\partial\mathcal{B}_z/\partial z$. Fits (dashed curves) to a simple lineshape model (see Methods) show good agreement with the data. $\delta = 0$ is defined as the point where all curves cross. The error bars in **b** and **c** represent 1σ statistical uncertainties.

Extended Data Table 1 | Parameters varied in the search for systematic errors

Category I Parameters
Magnetic fields
- $\mathcal{B}$-field gradients:
$\frac{\partial \mathcal{B}_z}{\partial z}, \frac{\partial \mathcal{B}_z}{\partial y}, \frac{\partial \mathcal{B}_x}{\partial x}, \frac{\partial \mathcal{B}_y}{\partial y}, \frac{\partial \mathcal{B}_y}{\partial x}, \frac{\partial \mathcal{B}_z}{\partial x}$ (even and odd under $\tilde{\mathcal{B}}$)
- Non-reversing $\mathcal{B}$-field: $\mathcal{B}_z^{\mathrm{nr}}$
- Transverse $\mathcal{B}$-fields: $\mathcal{B}_x, \mathcal{B}_y$ (even and odd under $\tilde{\mathcal{B}}$)
- $\tilde{\mathcal{E}}$-correlated $\mathcal{B}$-field: $\mathcal{B}_z^{\mathcal{E}}$
(to measure suppression of possible $\phi^{\mathcal{E}}$ effects by the $\tilde{\mathcal{N}}$ switch)
Electric fields
- Non-reversing $\mathcal{E}$-field: $\mathcal{E}^{\mathrm{nr}}$
- Field plate ground voltage offset
Laser detunings
- Detuning of refinement/readout lasers: Δ_{ref}, Δ_{read}
- 1-photon, 2-photon detuning of STIRAP lasers
- $\tilde{\mathcal{P}}$-correlated detuning: $\Delta^{\mathcal{P}}$
- $\tilde{\mathcal{N}}$-correlated detuning: $\Delta^{\mathcal{N}}$
- Detuning of rotational cooling lasers
Laser powers
- $\tilde{\mathcal{N}}\tilde{\mathcal{E}}$-correlated power, $P^{\mathcal{N}\mathcal{E}}$
- Power of refinement/readout lasers: $P_{\mathrm{prep}}, P_{\mathrm{read}}$
- $\tilde{\mathcal{N}}$-correlated power, $P^{\mathcal{N}}$
- $\tilde{\mathcal{P}}$-correlated power, $P^{\mathcal{P}}$
- Readout $\hat{X}$-,$\hat{Y}$-dependent laser power
Laser pointings/position along $\hat{x}$
- Pointing change of the refinement/readout lasers
- Readout $\hat{X}$-,$\hat{Y}$-dependent laser pointing
- Position of refinement beam along $\hat{x}$
Laser polarization
- Polarization rotation of readout laser
- Readout polarization dither angle, θ
- Refinement/readout laser ellipticity
Molecular beam clipping
- Clipping of the molecular beam along $\hat{y}$ and $\hat{z}$
(changes transverse velocity and position of the ensemble)
Category II Parameters
Experiment Timing
- Readout $\hat{X}, \hat{Y}$ polarization switching rate
- Allowed settling time between block switches
Analysis
- Signal size cuts, asymmetry magnitude cuts, contrast cuts
- Spatial dependence of fluorescence recorded by the eight PMTs
- Variation with time within the molecular pulse
- Variation with time within the $\hat{X}$,$\hat{Y}$ polarization cycle
- Search for correlations with all ω, C switch-parity components
- Search for correlations with auxiliary monitored parameters
($\mathcal{B}$-fields, laser powers and frequencies, vacuum pressure,
environment and beam source pressures and temperatures)
- 4 analyses of the data

Category I, parameters that we vary far from their typical value during data collection. We directly measure or place limits on the error, which leads to a linear shift in $\omega^{\mathcal{N}\mathcal{E}}$. Category II, parameters for which all values are consistent with normal conditions of the experiment. These served as checks for the presence of unexpected systematic errors.

ARTICLE

https://doi.org/10.1038/s41586-018-0591-3

Elucidating the control and development of skin patterning in cuttlefish

Sam Reiter[1], Philipp Hülsdunk[1,2], Theodosia Woo[1], Marcel A. Lauterbach[1], Jessica S. Eberle[1], Leyla Anne Akay[1], Amber Longo[1], Jakob Meier-Credo[1], Friedrich Kretschmer[1], Julian D. Langer[1,3], Matthias Kaschube[2] & Gilles Laurent[1]*

Few animals provide a readout that is as objective of their perceptual state as camouflaging cephalopods. Their skin display system includes an extensive array of pigment cells (chromatophores), each expandable by radial muscles controlled by motor neurons. If one could track the individual expansion states of the chromatophores, one would obtain a quantitative description—and potentially even a neural description by proxy—of the perceptual state of the animal in real time. Here we present the use of computational and analytical methods to achieve this in behaving animals, quantifying the states of tens of thousands of chromatophores at sixty frames per second, at single-cell resolution, and over weeks. We infer a statistical hierarchy of motor control, reveal an underlying low-dimensional structure to pattern dynamics and uncover rules that govern the development of skin patterns. This approach provides an objective description of complex perceptual behaviour, and a powerful means to uncover the organizational principles that underlie the function, dynamics and morphogenesis of neural systems.

Cuttlefish and octopuses have an unmatched ability to change their external appearance for camouflage or communication[1]. When camouflaging, they produce a statistical approximation of their visual environment following rules that remain unknown. Because cephalopod camouflage appeared evolutionarily as a response to predators and because their performance can fool humans as well, the rules of pattern generation that they express may be instructive to the texture perception across animals, and reveal biological solutions to a general problem of computational vision and neuroscience[2–6].

Since pioneering work on cephalopod chromatophores in the 1960s[7,8], several groups have revealed the remarkable complexity of this system[9–11]. Pigment-carrying chromatophores—the pixels of this two-dimensional texture generation system—expand and contract in direct response to the activity of motor neurons[8], which project from the brain[12] and make excitatory glutamatergic synaptic connections[13] with sets of muscles that are arranged radially[14]. Chromatophores operate in concert with other specialized cells (for example, leucophores and iridophores) and dermal muscular systems to generate a rich array of coordinated textures, dynamic patterns and behaviours.

The rules of neural control governing this system remain largely unknown, owing mostly to the challenges of tracking large numbers (thousands to millions) of small (15–100-μm diameter) chromatophores in soft-bodied, behaving animals. Because each chromatophore is controlled by a small number of motor neurons and conversely, because each motor neuron controls a small number of chromatophores (its motor unit)[12,14,15], we reasoned that chromatophore expansion could serve as a proxy for motor neuron activity. Analysis of the joint statistics of chromatophore variation might in turn reveal the structure of a hypothetical control hierarchy. This study presents our solutions to this challenge, a method for tracking nearly all chromatophores of the dorsal mantle of a cuttlefish at a high frame rate and over developmental timescales. Using this technique, we take the first quantitative steps towards elucidating the control, dynamics and morphogenesis of this system.

Tracking chromatophores in freely behaving animals

Freely behaving animals were filmed in a tank with variable backgrounds (Methods). Our first goal was to segment all visible chromatophores in all images and then align these images by mapping them into a common reference frame (Fig. 1a). In any recording session, continuous image sequences with the animal in view and in focus ('chunks', Fig. 1b) were interspaced by unusable periods of movement (grey, Fig. 1b). Chunks were selected post hoc using a statistic of focus (Fig. 1b and Methods). Within a chunk, the pixels of each frame were first classified[16] as belonging either to a chromatophore (of any colour) or to background (Fig. 1c, Extended Data Fig. 1 and Methods). All frames in one chunk were mapped into a common reference frame using sparse optical flow[17] (Fig. 1d and Supplementary Video 1) and averaged into one 'master frame' (Fig. 1e, top row, left and middle).

To track individual chromatophores across filming gaps, we stitched chunks together. By correlating a small patch of skin (purple and green frames, Fig. 1e) within one master frame with all possible positions and orientations of identically sized patches in another master frame, a single 'fingerprint' match was usually detected (see two-dimensional correlations in Fig. 1e, note the sensitivity to small shifts in the bottom left panel; see also Extended Data Fig. 2). We used a matching procedure over a grid of patches (Fig. 1e, top left and bottom right, red), interpolating between matching points, to map all master frames into a common reference frame (Methods and Supplementary Video 2). Over 970 analysed master frames, 85% had an average mapping error of $\leq$3 pixels per 20 ± 6 μm. The following data come from 1,178,146 mapped frames, from six animals.

Averaging all aligned master frames in a dataset produced a single 'queen frame'. Using local maxima, we partitioned the queen frame into non-overlapping sectors, each representing the space that a single chromatophore can occupy (Fig. 1f). The number of chromatophore pixels in each sector was then used to quantify the expansion state of the corresponding chromatophore in each image of the dataset (Fig. 1g). We thus obtained tens of thousands of parallel and simultaneous chromatophore–activity times series, characterizing skin patterns and their evolution (Methods).

[1]Max Planck Institute for Brain Research, Frankfurt am Main, Germany. [2]Frankfurt Institute for Advanced Studies and Department of Computer Science and Mathematics, Goethe University, Frankfurt am Main, Germany. [3]Max Planck Institute of Biophysics, Frankfurt am Main, Germany. *e-mail: gilles.laurent@brain.mpg.de

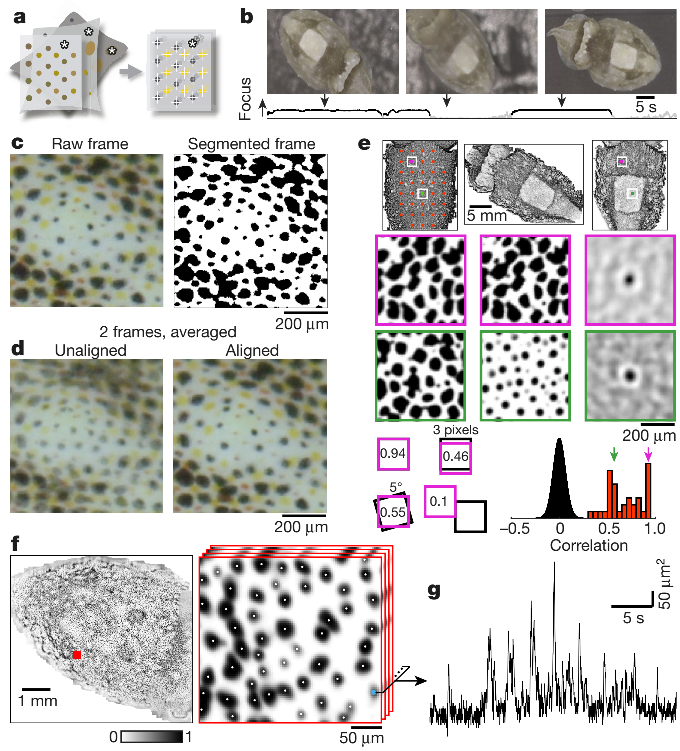


Fig. 1 | Chromatophore tracking in behaving cuttlefish. a, Schematic of approach to track the expansion state of many single, identified chromatophores over time through nonlinear image alignment. An example chromatophore is indicated by an asterisk. **b**, Data acquisition. 'Chunks' of in-focus image segments are identified with a focus statistic (black, bottom), separated by out-of-focus segments (grey). **c**, Segmentation. Pixels are classified as belonging either to a chromatophore (black, right) or to background (white). **d**, Alignment. Compare the averages of two frames with and without within-chunk alignment. **e**, Stitching. Top row, one master frame (middle) is mapped into the reference frame of another (left), resulting in the image on the right. Middle two rows, corresponding patches of skin (purple and green dots in the top images), after alignment (left and middle). Matches were made possible by the peak in spatial cross-correlation function (right) even when chromatophore states differ (green). Bottom row, left, shifting or rotating corresponding patches (purple frame) rapidly decreases correlation (Pearson's correlation). Bottom row, right, distribution of correlation values of 64 × 64-pixel skin patches (as above) across aligned master frames at all translations (sampled every pixel) and rotations (sampled every 60°). Correct matches (red): $n = 36$. Non-matching pairings (black), $n = 485{,}722{,}908$. **f**, Defining chromatophores. Left, queen frame ($n = 167{,}065$ aligned, averaged, segmented frames). Right, zoom-in of the red square on the left; single-chromatophore centres indicated with white dots. **g**, Raw expansion state of chromatophore in **f** (blue) over successive frames.

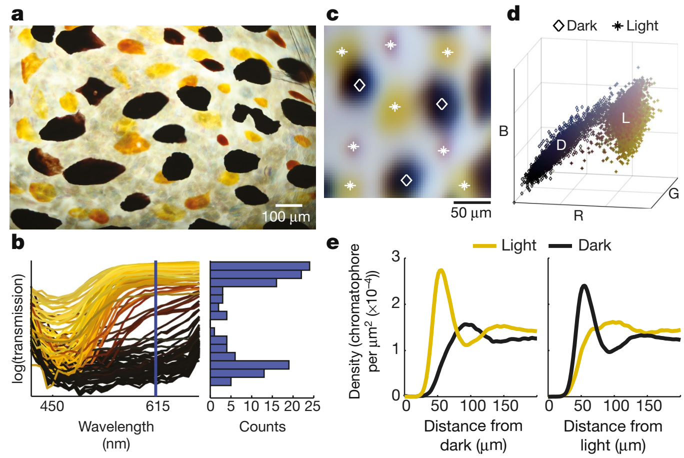


Fig. 2 | Classifying chromatophores by colour. a, Isolated skin sample illustrating range of chromatophore colours. **b**, Transmission spectra of chromatophores, including those shown in **a**. Lines correspond to regions of interest of individual chromatophores, coloured as the average RGB value for that region of interest. Right, frequency distribution histogram for transmission at 615 nm. **c**, Small patch of averaged, aligned colour images ($n = 28{,}998$ frames) showing dark (diamond) and light (star) colour assignments. **d**, Normalized RGB colour distribution for pixels centred on $n = 9{,}199$ chromatophores on the mantle of one animal. Symbols as in **c**. **e**, Radial-averaged density of light and dark chromatophores centred on dark (left) and light (right) reference chromatophores. $n = 11{,}535$ dark; 14,802 light chromatophores, from three animals.

Classifying chromatophores by colour

Cuttlefish chromatophores come in different colours (Fig. 2a), which are usually classified in 3–5 groups[9,18,19]. To characterize chromatophore colour objectively, we measured their transmission spectra in freshly dissected skin (Fig. 2b and Methods). The distribution of spectra at 615 nm was bimodal, with a 'dark' and a 'light' cluster, plus intermediate colours ranging from orange to red. This could be explained in part by expansion state: local application of L-glutamate[13]—which induces chromatophore expansion—caused spectral changes towards lighter colours (Extended Data Fig. 3), consistent with previous descriptions[20,21], and possibly explained by decreased local density and nano-structural features of the pigment granules[22]. Using mass spectrometry-based techniques, we identified xanthommatin as a pigment in *Sepia* skin, and localized it exclusively to light chromatophores (Extended Data Fig. 3). Therefore, we can segregate chromatophores of *Sepia officinalis* into two groups (light and dark) defined, respectively, by the presence and absence of xanthommatin. This confirms our initial classification based on transmission spectra (Fig. 2b).

Consistent with results in vitro, chromatophore colour in vivo defined two modes, with partially overlapping dark and light (yellow to brown) clusters (Fig. 2c, d and Methods). The spatial arrangement of chromatophore colour was not random (Fig. 2e): we calculated the average local density of each colour class centred on chromatophores of a single colour (Fig. 2e; 32,740 chromatophores, three animals; Methods). On average, chromatophores of either class occupied an approximately 20-µm radius area. Beyond this, the density of opposite-colour chromatophores increased and dominated, peaking at around 55 µm. At about 100 µm, colour densities were inverted, indicating an alternation (on average) between light and dark chromatophores (see also Extended Data Fig. 4).

Decomposing chromatophore control

To infer the potential structure of the control circuitry of the chromatophores, we examined their temporal co-variation during spontaneous fluctuations in vivo. This analysis is complicated by the fact that each chromatophore may be innervated multiple times and the possibility that individual motor units may overlap[14,23–28] (Extended Data Fig. 5).

To identify motor units directly, we first carried out minimal electrical stimulation of distal nerve branchlets that innervate freshly dissected dorsal mantle skin and measured resulting chromatophore expansion (Methods). Putative motor units were small (2–10 chromatophores), usually clustered (median radius of 2.5 chromatophores or 247 µm) and monochromatic (Fig. 3a and Extended Data Fig. 5), consistent with observations in squid and octopuses[12,21,27,28]. Light motor units were harder to stimulate electrically in isolation than dark ones, suggesting smaller axons.

Close observation of behaving animals revealed pronounced coordinated fluctuations of small localized groups of chromatophores, suggesting common drive (Supplementary Video 3). Seeking to extract these groups statistically, we factorized the chromatophore activity matrix using independent component analysis[29,30] and clustered through thresholding chromatophores that contributed strongly to single independent components, allowing for multiple-cluster membership (Methods). We called these inferred clusters of chromatophores 'motor elements' to distinguish them from anatomically defined motor units (see above).

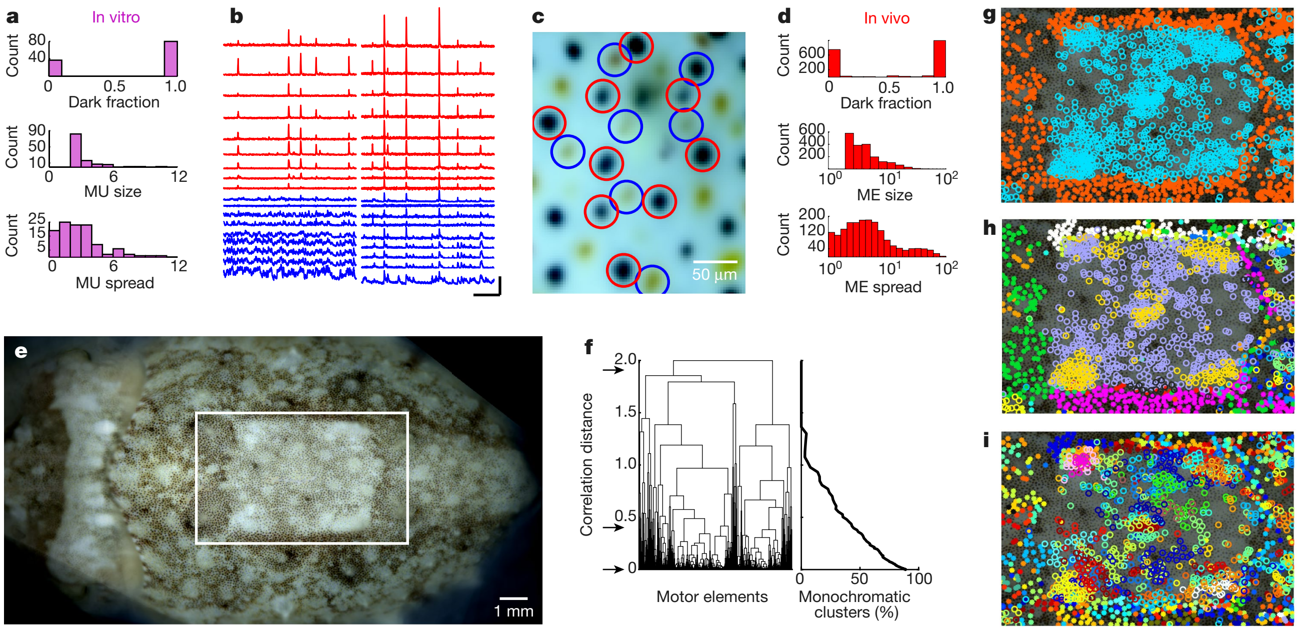


Fig. 3 | Inferring chromatophore neural control from co-variation. **a**, Summary statistics of in vitro experiments. Top, fraction of dark chromatophores in a motor unit (MU). Middle, number of chromatophores in a motor unit. Bottom, average distance to motor unit centroid for all chromatophores in that motor unit, normalized to the average nearest-neighbour distance over all chromatophores ($n = 295$ chromatophores, 114 motor units). **b**, Identification of motor elements (MEs) in vivo. Size-over-time traces for 18 chromatophores over two chunks. Red, nine chromatophores clustered as one motor element. Blue, nearest neighbours (in physical space) to the chromatophores in red. Calibrations: $x = 5$ s; $y = 1{,}000\ \mu m^2$. **c**, Average colour image showing position and colour of chromatophores in **b** (circles, colours). **d**, Summary statistics of in vivo experiments. Compare to **a**. **e**, Aligned colour image showing 'average' pattern over an approximately 1 h-long dataset (237,826 frames). **f**, Left, correlation-based hierarchical clustering of average motor element time courses for the dataset in **e** ($n = 695$ motor elements). Right, fraction of monochromatic clusters as a function of correlation distance ($n = 1{,}896$ motor elements, 3 animals). **g**–**i**, Clusters at threshold levels in **f** (arrows, top to bottom) within frame in **e**. Same symbols throughout; colours denote cluster identities.

Using 57 ± 10-min-long datasets, we extracted hundreds of motor elements across the mantle of an animal. Although chromatophores within a motor element tended to be highly correlated with each other, we often observed subsets within a motor element that occasionally fluctuated independently of the others. We also regularly saw transient co-fluctuations with unclustered, otherwise weakly correlated, chromatophores (Fig. 3b).

Motor elements were mostly monochromatic (89% contained only light or dark) and their size distribution was heavy-tailed, with a median of three chromatophores (Fig. 3d and Extended Data Fig. 5). Note that their size distribution resembles that of presumed motor units (Fig. 3a, d)—with the tail likely to represent groups of highly coordinated motor units, identified by independent component analysis. Chromatophores within a motor element were typically clustered physically, but were usually not 'nearest neighbours', consistent with colour alternation (Fig. 2e) and monochromaticity. The distribution of spatial clustering was also heavy-tailed, with few motor elements containing chromatophores spread over large areas (Fig. 3d).

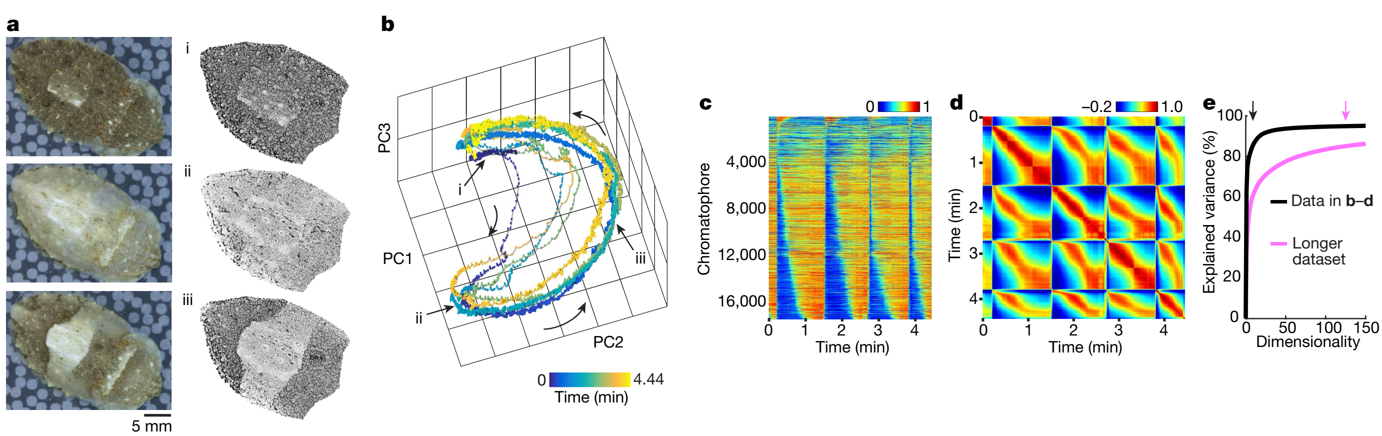


Fig. 4 | Tracking pattern changes at cellular resolution. **a**, Snapshots (i–iii) of an animal reacting to motion. Left, raw images. Right, corresponding segmented images: chromatophores are shown as disks proportional to actual size ($n = 17{,}305$ chromatophores, see Supplementary Video 5). **b**, Full sequence of skin patterns (17,305-dimensional vectors of chromatophore sizes, unfiltered) over repeats of the behaviour, projected into space of the first three principal components (PC1–PC3). Time is shown in colour. **c**, Full area-over-time matrix of the behavioural sequence in **b**. Chromatophore areas normalized for visualization (0–1) and ordered by time-to-cross mean activity during first sequence (15–88 s). **d**, Correlation matrix of full, 17,305-dimensional vectors of chromatophore sizes (Pearson's correlation). **e**, Cumulative variance explained by increasing numbers of principal components (dimensions). Black, sequence in **b**–**d**; magenta: 37-min dataset, including more patterns. Arrows indicate x where $y = 85\%$.

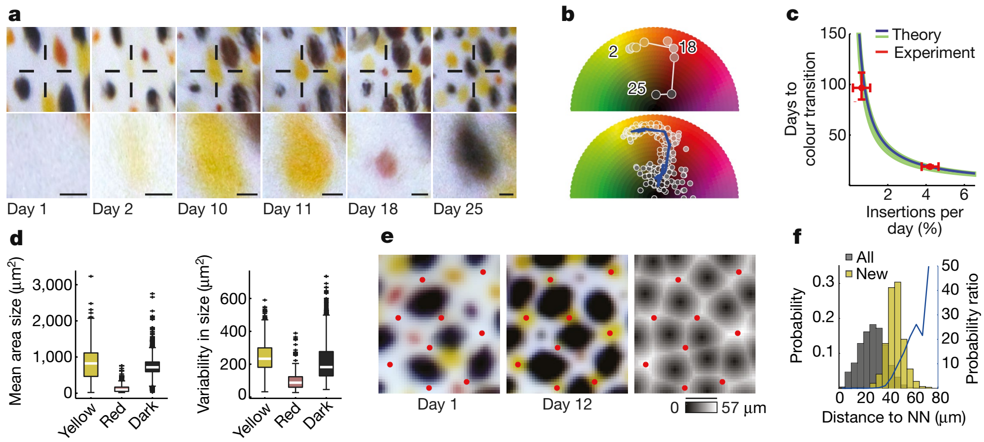


Fig. 5 | Tracking development of the chromatophore array. a, Aligned images of the same small skin patch over days, illustrating birth and colour changes. Crosshair zoomed-in below. 'Day 1' applied post hoc to day preceding detection. Scale bars, ~50 μm (nonlinear alignment). **b**, Top, colour evolution over days for chromatophore in **a**, plotted in hue–value space. Bottom, same for $n = 13$ chromatophores. After white-balancing, some black chromatophores appear bluish due to noise and are not shown (lower half of the colour wheel). Blue line, median colour over all chromatophores at same estimated developmental time ($n = 13$). **c**, Theoretical relationship between new chromatophore insertion rate and colour maturation (transition to black) for light (L) to dark (D) ratio measured in a juvenile over development (blue), and for the distribution of L/D ratios measured over 1–8-month-old animals (green shows ±1 s.d. from the mean, $n = 7$ animals, see Supplementary Information). Note that experimental measurements (red) fall precisely on the theoretical curve (horizontal error bars: s.e.m., from cross-validation). **d**, Distributions of size (left) and standard deviation of size over time (right) for yellow, red (transitional) and dark chromatophores. Transitional chromatophores are significantly smaller and less variable over time than either yellow or dark ones ($P = 3.5 \times 10^{-92}$ or 9.7×10^{-107}, respectively, Kruskal–Wallis test followed by Tukey's honest significant difference test; $n = 1{,}413$ yellow, $n = 214$ red, $n = 1{,}468$ dark). **e**, New chromatophores arise in gaps in existing array. i, ii, Same skin patch aligned 11 days apart. Red dots centred on chromatophores detected on day 12 but absent on day 1. iii, Greyscale shows distance to nearest older chromatophore in ii. **f**, Summary of insertion location statistics. Dark, distribution of distances to nearest old chromatophore. Yellow, the same distribution, conditioned on location of new chromatophore insertion. Blue line, probability ratio of yellow-to-black distributions, showing an approximately monotonic increase at increasing distances to nearest old chromatophore. $n = 11{,}527$ old; 1,412 new chromatophores, 2 animals. NN, nearest neighbour.

We next tested whether the statistical relationships between motor elements might reveal elements of higher-level control[9]. We averaged the time series of all chromatophores in single motor elements to approximate their underlying neural drive (Methods). We then performed hierarchical clustering on the correlation of these average time series, clustering chromatophores that formed motor elements at different levels of granularity.

We illustrate the approach with a dataset in which the animal displayed distinct macroscopic pattern components (Fig. 3e). The first bifurcation (Fig. 3f, top arrow) divided chromatophores within the white square and posterior spots[9] from all the others (Fig. 3g). Within these two superclusters, many degrees of correlation (that is, putative levels of common control) and monochromaticity could be identified. At lower levels in the hierarchy (Fig. 3f, middle arrow), the decomposition revealed medium-sized but still spatially structured elements, such as those that define the borders of the white square (Fig. 3h). Notably, some of these subpattern elements were at times observed to vary independently of their supercluster, consistent with previous brain stimulation experiments, indicating multiple levels of motor control[31] (Supplementary Video 4). In turn, these elements could be decomposed into motor elements (Fig. 3f, bottom arrow), often forming common or collinear patterns indicative of precise innervation, honouring the borders of macroscopic pattern elements (Fig. 3i).

Tracking pattern dynamics

Changes in the visual scene of an animal usually triggered rapid skin pattern changes. In the example in Fig. 4, a hand was moved above the cuttlefish (Fig. 4a), causing it to transition from dark to light (Fig. 4a and Supplementary Video 5). We examined this transition over several repeats by tracking the states of 17,305 chromatophores at 60 images per second (Fig. 4b–d). Projected into principal component space for visualization, the data took the form of looping trajectories joining dark (Fig. 4b, i) and light (Fig. 4b, ii) states through a sequence of intermediate states (for example, iii; Fig. 4b). Upon each stimulus, the animal not only generated the same target patterns (i and ii) but also moved with chromatophore-level repeatability (Extended Data Fig. 6) through very similar, low-dimensional sequences of intermediate states (Fig. 4c–e: 85% of the variance is explained by nine dimensions). This reliability of sequential chromatophore activation is remarkable because no physical constraints—such as those imposed by a moving limb for example—exist to prevent arbitrary and possibly more direct transitions. This suggests that neural constraints, probably linked to the putative control hierarchy inferred above (Fig. 3 and Extended Data Fig. 7) and to internal connectivity, limit the paths along which pattern transitions can occur.

Tracking array development

S. officinalis continuously add new chromatophores as they grow, increasing from a few thousands in hatchlings to a few millions before death[9]. To track chromatophore insertion and development, we aligned multiple datasets that were recorded days apart (Methods and Supplementary Video 6). We observed that all chromatophores change colour in a systematic progression: all newly born chromatophores were pale yellow (Fig. 5a and Extended Data Fig. 8), consistent with observations in hatchlings[32]. In a seven-day-old animal, yellow chromatophores transitioned over the course of around two weeks to orange and later, briefly, to red. Then, 18.7 ± 1.1 days after detection, each chromatophore turned dark and remained so throughout our observation period, possibly owing to xanthommatin polymerization (Fig. 5b and Extended Data Fig. 8; 13 chromatophores, 25 days). Therefore, the intermediate colours of chromatophores result from at least two causes: their expansion state (Extended Data Fig. 3) and their age (Fig. 5).

Chromatophores (1) do not disappear[33] and (2) their colour ratio (light/dark) is roughly constant (1.06 ± 0.19 to 1 in seven 8–252-day-old animals)[34]. However, (3) the time over which chromatophores turn from light to dark increased from around 19 days (above) to about 97 days in a 105-day-old animal (96.6 ± 9.3 days, mean ± s.e.m.; Methods). Likewise, whereas light chromatophores are produced daily as a fixed fraction of all existing chromatophores, (4) the rate of chromatophore addition dropped from 4.1% in a seven-day-old animal to

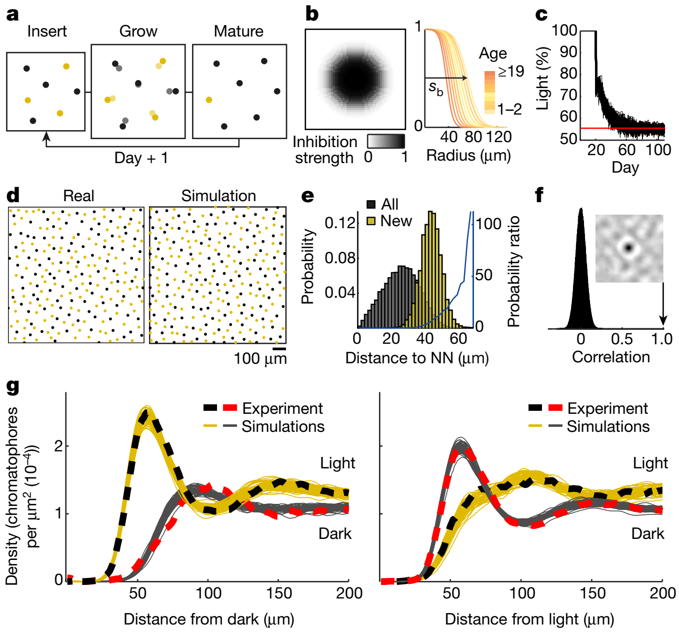


Fig. 6 | Simple rules explain the spatial layout of chromatophores. **a**, Schematic of model. Every day, new light chromatophores are inserted into the skin, chromatophores move apart as the skin grows, and chromatophores turn from light to dark when they reach a mature age (19 days). **b**, Left, disc of surround inhibition centred on each chromatophore. Right, radius of inhibitory surround decreases with age at rate r_a from an initial birth size s_b. **c**, Percentage of light chromatophores as a function of simulated day (50 runs). Red line, experimental ratio. **d**, Real versus simulated chromatophore position and colour assignment for small skin patches. **e**, Summary of insertion location statistics as in Fig. 5f, but for simulated developing skin. 'New' insertions are those on the last 'day' of simulation (4 out of 15,494 new insertions at 69–73 μm are not shown for clarity). **f**, Distribution of Pearson's correlation coefficients between a random patch and all locations of simulated skin. Inset: cross-correlation function centred on the correct matching location, corr. = 1. **g**, Radially averaged density of light and dark chromatophores from the centre of dark (left) and light (right) chromatophores. Dashed lines, experimentally measured densities ($n = 4{,}095$ dark; 5,104 light chromatophores). Thin lines, 50 independent runs of the model.

0.6% in a 105-day-old animal. In the Supplementary Information, we provide a formal derivation of an expression linking these observed quantities (2–4). Figure 5c shows the theoretical interdependence of two of these quantities (3 versus 4) given (2): the lifetime of the light state and insertion rate measured experimentally fall precisely on this curve, suggesting that these two properties are balanced to maintain a near-constant colour ratio across the life span of an animal.

The monochromaticity of motor units (Fig. 3a) could result from the fact that new motor neurons[9] innervate only newly born (light) chromatophores[21]. This hypothesis, however, introduces a conundrum in that each animal should keep track of the age of each motor unit to know its colour, an unlikely feat. This problem could be solved, however, if the chromatophores were re-innervated as they change colour, replacing 'light' with 'dark' motor neurons. Consistent with this, we observed that the average diameter and size variance over time of the red chromatophores were smaller than with the light and dark ones (Fig. 5d and Extended Data Fig. 9), suggesting that motor units undergo re-innervation as they are nearing transition to dark chromatophores.

We next examined the geometry of new-chromatophore insertions. The example in Fig. 5e (i, ii) shows the same aligned patch of skin with an interval of 11 days. The positions of the young (yellow) chromatophores at day 12 (red dots) retrospectively identified the zones of their future insertion as positions far from already born chromatophores (Fig. 5e (iii), 5f). This arrangement suggested a simple model for the development of the chromatophore array, based on the regulated production of a hypothetical inhibitory signal by each chromatophore[19,35], which we evaluated using computer simulations.

Simple rules can explain spatial layout

Our simulation ran on discrete steps (days) and was initiated by the random insertion of light chromatophores in a patch of bare skin, constrained by a chromatophore-centred inhibitory surround (Fig. 6a, b). Once filled with chromatophores, the skin patch 'grew' isotropically by a fixed proportion, followed by the next 'day' of chromatophore insertion. When chromatophores reached 19 days (above), they switched from light to dark (Fig. 6a). The inhibitory surround was described by a sigmoidal function derived from empirical measurements. To match the experimentally measured differences in spacing between newer and older chromatophores (Fig. 2e and Extended Data Fig. 9), we allowed the size of the inhibitory surround to change as chromatophores age. We fixed the shape of the surround, and fitted its initial size (s_b) as well as the rate of size change with age (r_a) to empirical measurements (Methods).

Consistent with our analytical results (Fig. 5c), this simple model converged to the observed percentage of light chromatophores (model = 0.55 ± 0.01, 50 simulations; data = 0.55, 5,104 light/4,095 dark, 1 animal; Fig. 6c), provided the skin growth rate was set to allow a realistic rate of chromatophore insertion (model = 4.23 ± 0.01%; data = 4.1% per day). It produced realistic spatial patterns of light and dark (Fig. 6d), new chromatophore insertion locations (Fig. 6e versus Fig. 5e) and chromatophore density (mean density of chromatophores per μm²: model $= 2.52 \times 10^{-4} \pm 0.1 \times 10^{-4}$; data $= 2.44 \times 10^{-4}$). Local patches of simulated skin had unique spatial layouts (fingerprints) of the type that we exploited for image registration (Fig. 6f versus Fig. 1e). Our model was able to produce realistic local chromatophore-centred densities, featuring the experimentally observed interdigitation of colour-specific modes (Fig. 6g versus Fig. 2e; Extended Data Fig. 9). Notably, by varying r_a, we could generate other known chromatophore distribution patterns such as the discoid units observed in some squid species[36] (Extended Data Fig. 10). This simple model may thus apply more generally to cephalopod skin patterning.

Discussion

We developed a strategy to track tens of thousands of individual chromatophores in freely behaving cephalopods, enabling studies of behaviour and development at cellular resolution. Our results open a path towards addressing many important biological questions. A first question concerns visual perception. Cephalopod camouflage is unique in revealing a high-dimensional neural readout of the visual texture perception of an animal. Identifying the primitives of cephalopod camouflage might not only reveal fundamental features of texture generation but also of vertebrate texture perception, because the former (in cephalopods) probably evolved in response to the latter (in their vertebrate predators). A second question concerns the development of methods to analyse very large neural datasets in the context of naturalistic behaviour[37]. Because chromatophore data can be assigned unambiguously to identified elements that lie at the same level of a neural hierarchy (here exclusively motor neurons), their analysis does not suffer from assumptions about their identities and positions in structured or recurrent circuits, as may happen with brain neural imaging. A third question concerns morphogenesis and development. Our data suggest that simple local rules can explain the structure of a continuously growing chromatophore array. They thus lead directly to clear questions about mechanisms and about their similarity with ones known from other systems[38–40]. Fourth, our results indicate that very complex behaviours can be described quantitatively at cellular resolution and in species that may reveal much about shared constraints on brain evolution[41]. This system is therefore particularly well-suited for studying the relationship between neural and behavioural dynamics, a central and general problem in neuroscience.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0591-3.

Received: 31 January 2018; Accepted: 8 August 2018;
Published online 17 October 18.

1. Messenger, J. B. Cephalopod chromatophores: neurobiology and natural history. *Biol. Rev. Camb. Philos. Soc*. **76**, 473–528 (2001).
2. Beck, J. & Gibson, J. J. The relation of apparent shape to apparent slant in the perception of objects. *J. Exp. Psychol*. **50**, 125–133 (1955).
3. Julesz, B. Visual pattern discrimination. *IRE Trans. Inf. Theory* **8**, 84–92 (1962).
4. Portilla, J. & Simoncelli, E. P. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis*. **40**, 49–70 (2000).
5. Gatys, L. A., Ecker, A. S. & Bethge, M. A neural algorithm of artistic style. Preprint at https://arxiv.org/abs/1508.06576v2 (2015).
6. Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V. & Shlens, J. Exploring the structure of a real-time, arbitrary neural artistic stylization network. Preprint at https://arxiv.org/abs/1705.06830v2 (2017).
7. Cloney, R. A. & Florey, E. Ultrastructure of cephalopod chromatophore organs. *Z. Zellforsch. Mikrosk. Anat*. **89**, 250–280 (1968).
8. Florey, E. & Kriebel, M. E. Electrical and mechanical responses of chromatophore muscle fibers of the squid, *Loligo opalescens*, to nerve stimulation and drugs. *Z. Vgl. Physiol*. **65**, 98–130 (1969).
9. Hanlon, R. T. & Messenger, J. B. Adaptive coloration in young cuttlefish (*Sepia officinalis* L.): the morphology and development of body patterns and their relation to behaviour. *Phil. Trans. R. Soc. B* **320**, 437–487 (1988).
10. Kelman, E. J., Osorio, D. & Baddeley, R. J. A review of cuttlefish camouflage and object recognition and evidence for depth perception. *J. Exp. Biol*. **211**, 1757–1763 (2008).
11. Gonzalez-Bellido, P. T., Scaros, A. T., Hanlon, R. T. & Wardill, T. J. Neural control of dynamic 3-dimensional skin papillae for cuttlefish camouflage. *iScience* **1**, 24–34 (2018).
12. Dubas, F., Hanlon, R. T., Ferguson, G. P. & Pinsker, H. M. Localization and stimulation of chromatophore motoneurones in the brain of the squid, *Lolliguncula brevis*. *J. Exp. Biol*. **121**, 1–25 (1986).
13. Florey, E., Dubas, F. & Hanlon, R. T. Evidence for L-glutamate as a transmitter substance of motoneurons innervating squid chromatophore muscles. *Comp. Biochem. Physiol. C* **82**, 259–268 (1985).
14. Reed, C. M. The ultrastructure and innervation of muscles controlling chromatophore expansion in the squid, *Loligo vulgaris*. *Cell Tissue Res*. **282**, 503–512 (1995).
15. Liddell, E. G. T. & Sherrington, C. S. Recruitment and some other features of reflex inhibition. *Proc. R. Soc. Lond. B* **97**, 488–518 (1925).
16. Ho, T. K. Random decision forests. In *Proc. 3rd International Conference on Document Analysis and Recognition* 278–282 (IEEE Computer Society, 1995).
17. Lucas, B. D. & Kanade, T. An iterative image registration technique with an application to stereo vision. In *Proc. 7th International Joint Conference on Artificial Intelligence* 674–679 (Morgan Kaufmann, 1981).
18. Mäthger, L. M., Chiao, C. C., Barbosa, A. & Hanlon, R. T. Color matching on natural substrates in cuttlefish, *Sepia officinalis*. *J. Comp. Physiol. A* **194**, 577–585 (2008).
19. Bassaglia, Y. et al. *Sepia officinalis*: a new biological model for eco-evo-devo studies. *J. Exp. Mar. Biol. Ecol*. **447**, 4–13 (2013).
20. Fioroni, P. Die embryonale Genese der Chromatophoren bei *Octopus vulgaris* Lam. *Acta Anat*. **75**, 199–224 (1970).
21. Packard, A. Morphogenesis of chromatophore patterns in cephalopods: are morphological and physiological 'units' the same? *Malacologia* **23**, 193–201 (1982).
22. Deravi, L. F. et al. The structure–function relationships of a natural nanoscale photonic device in cuttlefish chromatophores. *J. R. Soc. Interface* **11**, 20130942 (2014).
23. Dubas, F. Innervation of chromatophore muscle fibers in the octopus *Eledone cirrhosa*. *Cell Tissue Res*. **248**, 675–682 (1987).
24. Ferguson, G. P., Martini, F. M. & Pinsker, H. M. Chromatophore motor fields in the squid, *Lolliguncula brevis*. *J. Exp. Biol*. **134**, 281–295 (1988).
25. Messenger, J., Cornwell, C. & Reed, C. L-glutamate and serotonin are endogenous in squid chromatophore nerves. *J. Exp. Biol*. **200**, 3043–3054 (1997).
26. Maynard, D. M. in *Invertebrate Nervous Systems: Their Significance for Mammalian Neurophysiology* (ed. Wiersma, C. A. G.) 231–255 (Univ. Chicago Press, Chicago, 1967).
27. Dubas, F. & Boyle, P. R. Chromatophore motor units in *Eledone cirrhosa* (Cephalopoda: Octopoda). *J. Exp. Biol*. **117**, 415–431 (1985).
28. Florey, E. Nervous control and spontaneous activity of the chromatophores of a cephalopod, *Loligo opalescens*. *Comp. Biochem. Physiol*. **18**, 305–324 (1966).
29. Bell, A. J. & Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput*. **7**, 1129–1159 (1995).
30. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Netw*. **13**, 411–430 (2000).
31. Boycott, B. B. The functional organization of the brain of the cuttlefish *Sepia officinalis*. *Proc. R. Soc. Lond. B* **153**, 503–534 (1961).
32. Andouche, A. & Bassaglia, Y. Coleoid cephalopod color patterns: adult skin structures and their emergence during development in *Sepia officinalis*. *Vie Milieu* **66**, 43–55 (2016).
33. Packard, A. Size and distribution of chromatophores during post-embryonic development in cephalopods. *Vie Milieu* **35**, 285–298 (1985).
34. Yacob, J. et al. Principles underlying chromatophore addition during maturation in the European cuttlefish, *Sepia officinalis*. *J. Exp. Biol*. **214**, 3423–3432 (2011).
35. Packard, A. & Hochberg, F. G. Skin patterning in *Octopus* and other genera. *Symp. Zool. Soc. Lond*. **38**, 191–231 (1977).
36. Hanlon, R. T. The functional organization of chromatophores and iridescent cells in the body patterning of *Loligo plei* (Cephalopoda: Myopsida). *Malacologia* **23**, 89–119 (1982).
37. Gomez-Marin, A., Paton, J. J., Kampff, A. R., Costa, R. M. & Mainen, Z. F. Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nat. Neurosci*. **17**, 1455–1462 (2014).
38. Yamaguchi, M., Yoshimoto, E. & Kondo, S. Pattern regulation in the stripe of zebrafish suggests an underlying dynamic and autonomous mechanism. *Proc. Natl Acad. Sci. USA* **104**, 4790–4793 (2007).
39. Cheng, C. W. et al. Predicting the spatiotemporal dynamics of hair follicle patterns in the developing mouse. *Proc. Natl Acad. Sci. USA* **111**, 2596–2601 (2014).
40. Manukyan, L., Montandon, S. A., Fofonjka, A., Smirnov, S. & Milinkovitch, M. C. A living mesoscopic cellular automaton made of skin scales. *Nature* **544**, 173–179 (2017).
41. Kröger, B., Vinther, J. & Fuchs, D. Cephalopod origin and evolution: a congruent picture emerging from fossils, development and molecules: extant cephalopods are younger than previously realised and were under major selection to become agile, shell-less predators. *BioEssays* **33**, 602–613 (2011).

Acknowledgements We thank F. Bayer, A. Umminger and N. Heller for assistance in building the experimental setup; L. Jürgens, T. Klappich, J. Nenninger, E. Northrup and G. Wexel for animal care; R. Siegel for providing the squid skin image; E. Lamo Peitz, S. Trägenap and J. Racky for help with image alignment and segmentation; R. Hanlon and members of the Marine Biology Laboratory in Woods Hole for their hospitality to G.L. in the summer of 2016; Bruker Daltonics, J. Fuchser and C. Henkel for Fourier-transform ion cyclotron resonance–mass spectrometry measurement time and discussions; M. Tosches and L. Fenk for comments on the manuscript; and the members of the Laurent laboratory for discussions. Funded by the Max Planck Society (G.L.), the European Research Council (G.L.) and the Bernstein Focus: Neurotechnology Frankfurt (M.K.).

Reviewer information *Nature* thanks C. Machens, D. Osorio and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions G.L. conceived, initiated and managed the project. S.R., T.W., M.A.L., J.S.E., L.A.A., A.L. and G.L. designed and conducted the experiments. P.H., S.R., F.K. and M.K. developed and implemented the image-processing pipeline. J.M.-C. and J.D.L. carried out the mass-spectrometry analysis. M.A.L. developed the analytical model of colour evolution with input from M.K. S.R. developed and ran the numerical simulations with input from M.K. S.R., P.H., M.A.L., T.W., J.S.E., M.K. and G.L. analysed and discussed all data. G.L. and S.R. wrote the text with input from all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0591-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0591-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to G.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Experimental animals. Animal experimentation in this study was performed according to German animal welfare law (paragraph 11, sentence 1, #1, German animal welfare law to house and breed cephalopods for scientific purposes). European cuttlefish *S. officinalis* were hatched from eggs collected in the English Channel and reared in a seawater system, at 20 °C. The closed system contains 4,000 l of artificial seawater (ASW; Instant Ocean) with a salinity of 33‰ and pH of 8–8.5. Water quality was tested weekly and adjusted as necessary. Trace elements and amino acids were supplied weekly. Marine LED lights above each tank provided a 12/12-h light/dark cycle with gradual on- and off-sets at 07:00 and 19:00. The animals were fed live food (either *Hemimysis* spp. or small *Palaemonetes* spp.) ad libitum three times per day. Experimental animals of unknown sex were selected for healthy appearance and calm behaviour. The animals were housed together in 120-l glass tanks with a constant water through-flow resulting in five complete water exchanges per hour. Enrichment consisted of natural fine-grained sand substrate and seaweed (*Caulerpa prolifera*).

In vivo behavioural data acquisition. For in vivo behavioural experiments, six animals (1 to 60 days post-hatching, around 6–50 mm in mantle length) were placed in a capped filming chamber (150 mm × 95 mm × 75 mm or 240 mm × 170 mm × 50 mm) filled with seawater. A single filming session typically lasted between 10 and 90 min per day and per animal. Our filming procedures induced no pain, suffering, distress or harm to the animal. Naturalistic textures with normalized power spectra (Normalized Brodatz Texture database) and artificial patterns generated in MATLAB and Paint were displayed on the floor of the tank using an E-Ink display. Filming was performed at 59.94 frames per second (f.p.s.) in 4K full-frame (4,096 × 2,160) using the Sony PMW-F55 camera in the Sony RAW format. Resolution was $40.8 \pm 32.4\ \mu m^2$ per pixel. The camera was mounted on a motorized *x*–*y* translation stage and its position adjusted with a joystick to keep the animal in view. Acquired data were colour matched in DaVinci Resolve Studio 12.5 (Black Magic Design). Movies were compressed offline to the H.264 format using the x264 encoder provided by FFmpeg-2.8.6, with the compression preset 'faster' and the constant rate factor of 16, without chroma subsampling. These compressed movies were used for all subsequent in vivo data analysis. Note that the analyses and results presented in this paper do not depend on the exact statistics of the images or patterns shown to the animals. They depend only on our ability to detect changes and correlations between patterns produced by single animals over time, at sub-chromatophore resolution.

In vitro electrophysiology. For in vitro experiments, animals (19–40 weeks old, 115 ± 28 mm in mantle length, 10 in total) were euthanized according to well-established best-practice protocols[42]: animals were deeply anaesthetized first in isotonic 3% ethanol in ASW and then in 5% ethanol in ASW or using 3.5% (w/v) $MgCl_2$ in ASW. Superficial skin samples were then removed gently from the dorsal mantle, peeling away the superficial skin layers from the underlying body musculature, gently cutting nerves and connective tissue with iridectomy scissors and taking care that the chromatophores were not overly stretched or damaged. These skin samples were placed in cold ASW inside a transparent observation chamber, pinned at their edges, superficial surface down, stretched gently so as to eliminate wrinkles and left to recover. The chamber was placed on the translation stage of an inverted microscope and the chromatophore array observed with 1.25× or 2.5× objectives (126.75 or 291.78 pixels per mm). A fine suction electrode operated with a micromanipulator was placed on the cut end of a nerve and for electrical stimulation (pulse duration: 100 μs) at pulse rates of 0.5 Hz and at threshold intensity using a pulse generator (from ±5.57 to ±2,774 μA dialled up gradually, A.M.P.I., Master-8-cp) and a constant-current stimulus isolator (World Precision Instruments, A360). Colour images of the chromatophores were acquired with a CCD/CMOS camera (Basler acA1920-155uc) at 30 f.p.s. The stimulus trace and camera exposure times were recorded using a digitizer (Axon Digidata 1440A). Synchronization and analysis were conducted offline.

Transmission spectra. Transmission spectra were recorded from fresh skin samples (extracted as above). Samples were mounted on standard microscope slides with #1.5 coverslips in ASW. Transmission spectra were recorded in 32 channels on a Zeiss LSM 880 Examiner confocal microscope (10×, NA 0.45, water immersion objective) using the 'lambda mode' spectral detection. This mode is usually intended for fluorescence detection, but the halogen lamp for transmitted-light mode can be turned on with a service macro. Images were thus recorded with a scanned point detector but with wide field illumination. We manually drew regions of interest around chromatophores. Raw spectra were normalized with respect to a nearby region in the same field of view that did not contain any chromatophores. To measure the effect of expansion state on transmission spectra, ASW was replaced by a glutamate solution (40 μM) in ASW. Images were acquired before and 6–8 min after glutamate application.

Mass spectrometry. Excised skin tissue samples were homogenized in 1:1 methanol:water (v/v) supplemented with 1% trifluoroacetic acid, sonicated for 10 min in an ultrasonic bath and placed for 2 h on a rotary shaker. The samples were then centrifuged for 10 min, the supernatant removed, filtered through a 0.2-μm syringe filter and evaporated to dryness. For mass spectrometry, the dried extracts were resuspended in 95:5 water:acetonitrile (v/v) supplemented with 0.1% formic acid.

High-performance liquid chromatography coupled to ultraviolet light (UV) absorption and mass spectrometric detection (HPLC–UV–MS) experiments were carried out on an Ultimate 3000 RSLC system (Dionex) equipped with a CSH C18 column (Charged Surface Hybrid, 2.1 × 100 mm, 1.7-μm particle size, Waters) and variable wavelength detector set to 250 nm, coupled to an Impact II mass spectrometer (Bruker Daltonik). Separation was carried out using water (A) and acetonitrile (B), both supplemented with 0.1% formic acid, as mobile phases with a flow rate of $300\ \mu l\ min^{-1}$ at 40 °C. After 2 min of equilibration with 2% B, a linear gradient was ramped from 2% to 95% B in 30 min followed by 5-min wash (95% B) and 3-min equilibration (2% B) steps. The mass spectrometer was operated in positive-ion mode with a mass range of m/z 50–1,000. Processing and data analysis were performed manually using DataAnalysis 4.4 (build 200.55.2969, Bruker Daltonik).

For direct infusion experiments, extracts were diluted 1:100 and infused at $120\ \mu l\ h^{-1}$ into a 7T SolariX XR mass spectrometer (Bruker Daltonik). Spectra were recorded in positive-ion mode of m/z 107.5–2,000. For exact mass determination and fragmentation experiments, precursor ions were isolated using the quadrupole, inspected for contaminating ions and then subjected to collision-induced dissociation in the collision cell. Spectra were analysed with DataAnalysis 4.4 (build 200.55.2969, Bruker Daltonik).

For mass spectrometry imaging experiments, excised skin tissue samples were stretched and pinned onto frozen gelatine blocks, snap-frozen in isopentane and sectioned to 12 μm using a CM 3050s cryotome (Leica Biosystems). The slices were carefully transferred onto conductive ITO-coated glass slides (Bruker Daltonik), thaw-mounted and dried in a vacuum desiccator before taking optical slide scans with an OpticLab H850 histology slide scanner (Plustek). Samples were screened using a Rapiflex TOF/TOF mass spectrometer (Bruker Daltonik) operated in positive- and negative-ion modes, using a mass range of m/z 100–2,000. Ultrahigh-resolution mass spectra were acquired on a 7T SolariX XR mass spectrometer (Bruker Daltonik) in positive-ion mode in a mass range of m/z 107.5–2,000 using a 20 μm × 20 μm pixel grid. The laser was operated at 500 Hz with 100 shots per pixel and focus set to minimum. Imaging data were acquired and pre-processed using flexImaging 5.0 (build 5.0.78.0_1031_152, Bruker Daltonik) and further analysed and visualized using SCiLS Lab 2016b (build 4.01.8758, SCiLS). Individual images were adjusted to the same intensity scale and weak spatial denoising was applied for merged compounds. Spatial segmentation was performed with weak spatial denoising and a bisecting *k*-means algorithm based on the correlation distance of individual spectra. The relationship between colour and xanthommatin concentration was examined by manually clustering partitions of the *k*-means algorithm, corresponding to yellow and red–brown chromatophores.

Summary of the image processing and tracking pipeline. The major steps of the processing pipeline were as follows: (1) chunking: identify episodes of video ('chunks') with cuttlefish in focus; (2) segmentation: label pixels as chromatophore or background on individual frames; (3) registration: alignment across frames within chunks to correct nonlinear body distortions (over seconds); (4) stitching: alignment across chunks (seconds to hours); (5) chromatophore identification and size tracking; (6) colour assignment; and (7) stitching across days.

Chunking of in-focus frames. In vivo behavioural datasets consisted of series of frames in which the animal was in view and in focus, separated by frames in which the animal was out of view, out of focus or blurred owing to fast motion. We first identified in-focus frames using a simple focus statistic (sum of a difference-of-Gaussians filter size 15 × 15 pixels, $\sigma = 1.5$ and 2 pixels) to each image. The standard deviation of the filter was selected to match the mean size of chromatophores. Our statistics therefore indicated whether chromatophores were present and clear in an image. Continuous sequences of images were then selected semi-automatically, based on the amplitude of the focus statistic and its variability over images and time. We called continuous in-focus image sequences obtained within a single filming session 'chunks'.

Chromatophore segmentation. We segmented chromatophores from the background using a supervised learning approach. For training and validation, images (1 each) of 256 × 256 pixels containing a representative sampling of chromatophore sizes and colours were manually annotated pixel-wise as belonging to a chromatophore or background skin. Annotations were performed by six individuals and inconsistencies were removed using majority vote. We then fitted a random-forest model[16] to this annotation. Our model classified pixels as chromatophore or background based on feature vectors calculated from the output of eight difference-of-Gaussian filters ($\sigma = 0.8^{1.4x}$, $x = 1{:}8$) per RGB colour channel. Filter sizes were chosen to cover the range of observed chromatophore sizes. We determined the random forest parameters by hyperparameter optimization[43] (number of trees in 1–32, maximal depth in 1–32, minimal data size for split in 1–11, minimal data size for leaves 1–21, splitting criterion either by Gini impurity or information gain and

enabling or disabling bootstrap aggregation). Then, 1,000 models were fitted using a fourfold cross-validation and the best model was identified by the F_1 score. To this end, we used a model with eight trees with a depth of 8 and an entropy-based splitting criterion on five randomly selected features. We assessed model performance by comparing classifier performance against a second, manually annotated image (Extended Data Fig. 1).

Alignment of images within a chunk. Animal movement (for example, breathing) and skin distortions caused the pixel location of chromatophores to change over frames. Our high frame rate combined with the definition of chunking meant that differences in chromatophore locations (both affine and non-affine deformations) between successive frames within a chunk could be assumed to be small. We could therefore use image registration methods for small-displacement optic flow. We used the Lukas–Kanade optical flow algorithm[17] to track points centred on a random subset of around 300 round chromatophores. Round chromatophores were found by placing a threshold on circularity of chromatophores detected in the first frame of the chunk. These chromatophores were selected to minimize runtime. The full-frame optical flow was interpolated from these tracking points using a moving least-squares algorithm[44]. We chose a smoothness parameter ($\alpha = 3$) for interpolation to remove skin distortions and large movement, but not the fine scale movement of individual chromatophores.

Stitching averaged aligned images over chunks. Once all the images within a chunk were aligned, we averaged over the binarized images, generating a 'master frame'. The value of each pixel in a master frame thus represents the fraction of frames within the corresponding chunk in which that pixel was labelled as belonging to a chromatophore. Because chromatophore size can vary over frames during a single chunk, the typical profile of a chromatophore in a master frame is a radial gradient. After obtaining one master frame per chunk, we developed a method to register all of the chunks of a filming session into a common reference frame. We call this process 'stitching'. Individual chunks were, by definition, separated from each other by out-of-focus epochs, that is, frames in which the cuttlefish often changed position, angles in x, y and z, body shape and chromatophore pattern. Stitching thus required aligning and morphing chunks into the same reference frame. For every master frame in a dataset, we first defined a mask outlining the cuttlefish by applying a difference-of-Gaussian filter to the image and thresholding the result. We then mapped all master frames into each other's reference frames.

To stitch together two master frames 'A' and 'B', we first performed a coarse rigid-body transform mapping A into the reference frame of B by fitting an ellipse around the cuttlefish mask in both frames. This created B′, that is, B mapped into the reference frame of A through the inverse of this mapping. Next, we defined a grid of points 256×256 pixels apart over the cuttlefish mask of A. We attempted to find each point of this grid in B′ by correlating patches of 64×64 pixels centred on the grid points in A with regions in B′. We sampled a range of translations (± 256 pixels in 2-pixel steps) and rotations ($\pm 20°$ in 2° steps) around the pixel location of each grid point to find the pixel with the highest correlation value. In general, not all of the grid points could be correctly mapped; outliers were removed using the RANSAC algorithm[45] under an affine model. A new map was constructed from the remaining mapped points using moving least squares interpolation[44].

By applying the inverse of this mapping to B′ we produced B″, a more refined mapping. Fine alignment was performed by repeating this process on a finer grid. A new grid of points 32×32 pixels was defined on the cuttlefish mask of A. We then attempted to find each of the points in B″ with the highest local cross-correlation to 64×64-pixel patches centred on each grid point. We then interpolated between these points using moving least squares to produce a full map, B‴. Combining these three maps resulted in a single non-affine mapping from B into the reference frame of A.

This stitching algorithm was used to map every master frame in a dataset into the reference frame of every other master frame. We could quantify the accuracy of this non-symmetric mapping over the cuttlefish mantle by calculating the reprojection error: all points in the cuttlefish mask of master frame A were mapped into the reference frame of master frame B (using the A-to-B map) and then mapped back into the reference frame of A (using the B-to-A map). The reprojection error was defined as the Euclidian distance between the original and remapped points. A point was considered well-mapped if it reprojected to within three pixels (20 ± 6 μm) of its original location. By taking the fraction of well-mapped points in every master frame, we produced a matrix quantifying how well every master frame mapped into every master frame over the cuttlefish mantle. The column of this matrix with the highest well-mapped fraction identified the master frame into which all others mapped best. We used this as common reference frame for the dataset (see 'Chromatophore definition'). Poorly mapped chunks of data, defined as master frames that had less than 50% well-mapped points within the cuttlefish mask were excluded from subsequent analysis. Stitching failures were usually the result of poorly registered chunks, resulting in blurry master frames. These failures, in turn, were often due to temporary loss of focus through cuttlefish moving in and out of the focus range of our optical system.

Chromatophore definition. As done for within-chunk alignment, we used the maps generated by our stitching algorithm to project all master frames into the common reference frame of each dataset. The resulting average frame (the 'queen frame') represented approximately the probability of each pixel being labelled as belonging to a chromatophore throughout all in-focus frames over the entire filming session, excluding poorly mapped chunks of data (see 'Stitching averaged aligned images over chunks'). Chromatophores were detected by finding local maxima using a 3×3 footprint kernel. Applying the watershed transform to the queen frame with detected chromatophores as markers divided the image into basins, each defining a region surrounding a single chromatophore. The watershed transform also split groups of merged chromatophores, relying on gradients in the queen frame created by chromatophore size changes over the dataset. A conservative mask defining the region of the cuttlefish mantle that was in focus was drawn as a convex hull of manually selected points, and all subsequent analysis was performed on chromatophore basins within this masked region.

Chromatophore-size tracking. With the chromatophore basins so defined, we could track the expansion state of each chromatophore over time. Each segmented image (see 'Chromatophore segmentation') was mapped into the reference frame of the first image in a chunk (see 'Alignment of images within a chunk'), and then mapped again into the common reference frame of the dataset (see 'Stitching average aligned images over chunks'). The number of pixels classified as belonging to chromatophores in the segmented image was counted in every watershed basin of the queen frame. This pixel number, multiplied by an experiment-specific μm per pixel calibration, defined the area of each chromatophore in each frame. This calculation was repeated with all images in a dataset, producing parallel time-series measurements of size over time for all segmented chromatophores. Although chromatophore size generally varies with mantle position as the animal adopts different skin patterns, a uniformly dark pattern, as seen in Fig. 4a, reveals no strong correlation between chromatophore size and either anterior–posterior or medial–lateral position ($r = 0.06$ or $r = -0.03$, respectively).

Colour assignment. Determining the colour of chromatophores is difficult to accomplish accurately on single images in vivo owing to camera pixel noise, variability in lighting and the expansion-state dependency of chromatophore colour. We therefore analysed chromatophore colours by mapping all images of a dataset into the common reference frame of the dataset, producing an average colour image. We first constructed a feature space in which chromatophores could be accurately colour-labelled independently of our segmentation algorithm. We high-pass filtered the image and then performed independent component analysis in colour space[29,30]. After projecting the image onto the two largest independent components and thresholding each projection separately, we took the maximum value over projections. This image was smoothed with a Gaussian filter (s.d. = 1 pixel). We then applied a watershed algorithm[46] to identify chromatophore regions. Chromatophore centres were defined as the weighted centroid of each region. Visible chromatophores that were not detected automatically, typically smaller red and yellow chromatophores, were identified manually. The average RGB value of a region of 3×3 pixels of the average colour image, centred on the location of each chromatophore, defined chromatophore colour. These colours were clustered into two classes by fitting a Gaussian mixture model. We observed no strong correlation between the anterior–posterior or medial–lateral position on the mantle and chromatophore colour label ($r = -0.0047$ or $r = -0.0029$). We defined the colour of tracked chromatophores by performing a nearest-neighbour matching between chromatophore centres defined using our tracking pipeline and centres defined on the average colour image. For the motor element inference experiments, 93% of dark and 82% of light chromatophores that were detected in the average colour image ($n = 39{,}948$) could be linked to a tracked chromatophore ($n = 35{,}062$) located within three pixels (15.1 ± 5.9 μm, from three animals).

Stitching over filming sessions and detecting new chromatophores. We filmed two additional animals over periods of several weeks, that is, periods over which the animals underwent considerable growth. To track chromatophores over days and weeks, we used a modified version of our stitching algorithm (see 'Stitching average aligned images over chunks'). To model cuttlefish growth, we used a similarity transformation rather than a rigid-body transformation in the initial coarse alignment of cuttlefish masks. In the subsequent two correlation-matching steps, we use a larger search space in scales, rotations and translations. We used the resulting map to warp the queen frame of one dataset into the reference frame of another dataset. After this, we linked chromatophores over days through a nearest-neighbour matching. A convex hull containing the intersection of the chromatophore basins from both datasets was first calculated, and matching was only performed within this region. New chromatophores were defined as chromatophores from the later dataset that were located within the convex hull intersection and were not matched with a chromatophore in the earlier dataset. In total, $75 \pm 14\%$ of the pixels in the convex hull that contained mapped

chromatophores had reprojection errors below 50 μm (six datasets mapped from three animals), allowing for unambiguous matching through manual inspection and modification of matches using a custom graphical user interface.

Pipeline implementation. Our alignment and tracking pipeline was implemented on two computing clusters: the Draco supercomputer at the Max Planck Computing and Data Facility, where 16 jobs were processed in parallel on 1–2 nodes with 32 cores at 2.3 GHz (128 Gb RAM per node); and the FIAS computing cluster, where 3 jobs were processed in parallel on 2–4 nodes with 32–64 cores per node at 2.6 GHz (64 Gb RAM per node). Data management between local storage and compute nodes was managed by a Bash script determining the sending and receiving of data and configuration files, and starting the pipeline on compute nodes. On each compute node, the pipeline computations were managed by a second Bash script, which inserted all pipeline steps into a SLURM[47] queue. Parallel computation of the steps was handled by the SLURM controller. Parallelization per step was achieved by spawning programs using MPI[48], and distributing computations across program instances. For steps for which random access across video frames could not be implemented (for example, reading a video), MPI spawned programs following a one-producer/multiple-consumers pattern. For registration, parallelization was achieved by distributing the optical flow and moving least-squares algorithms across program instances. Threading was done using third-party libraries. The pipeline was written for GNU/Linux operating systems in the Python and Cython programming language, relying on Scipy[49], scikit-learn[50], scikit-image[51], PyQt and OpenCV-Python. Further data analysis was performed in Python and MATLAB. The results of all pipeline steps were stored using the HDF5 format. The pipeline was constructed so that each step read one file and output another. In total, we achieved an overall speed of around 1–1.5 frames s^{-1} (corresponding to around 2.4–3.6 Mb s^{-1} of a compressed video).

Chromatophore-triggered density plots. We constructed images composed of the locations of certain chromatophores (light, dark and so on) and then averaged regions of these images centred on the location of chromatophore classes of interest. The resulting chromatophore-triggered average image was linearly interpolated to 1 pixel per μm^2 and then smoothed with a Gaussian filter (15 μm size, 2–3 μm s.d.) We then computed radial averages.

Automated extraction of motor units (in vitro experiments). Consecutive frames were first aligned using optic flow (see 'Alignment of images within a chunk') to correct for spontaneous skin movements. For experiments containing >7 consecutive stimulus trials, the nearest video frames between 10-ms pre- and 200-ms post-stimulus were inspected and annotated manually for expansion events using image subtraction. Chromatophore position and colour were determined from the pre-stimulus frame of the first expanding trial. Colour classification was made using a threshold on the red channel of the white-balanced, contrast-stretched RGB space. Motor units were identified by coincident responses and failures of a set of multiple chromatophores (>1) across trials, allowing for an individual failure rate of up to 25%. We estimated the average spontaneous expansion probability (of 0.0293) by examining the activity of 81 chromatophores from four animals at times without stimulation. We then could estimate the chance probability of an observed sequence of responses and failures as

$$\binom{n}{e}(r^m)^e((1-r)^m)^{n-e}$$

in which m is the number of chromatophores in a putative motor unit, e is the number of expansion trials and n is the total number of trials. A threshold of $P=0.05$ was placed on observed sequences for inclusion as a motor unit. Motor units along the edge (mean + 1 s.d. of the average chromatophore nearest-neighbour distance) of the field of view were excluded from the analysis to prevent underestimation of motor unit size.

Inference of motor elements from in vivo imaging data. Our choice of statistical model for motor unit inference was motivated by the desire to capture the potentially overlapping innervation of motor units while excluding sets of chromatophores that are more transiently coordinated.

Chromatotophore–area time series were symmetrically low-pass filtered to 4 Hz using a 3-pole Butterworth filter. They were then downsampled 4–8-fold. We performed a 'spatial' ICA on the resulting matrix, using the Fast ICA algorithm[30]. This algorithm iteratively estimates $S=WX$, in which X is the centred, whitened, area-traces × chromatophore matrix, W is the unmixing matrix, and S is the component × chromatophore matrix of independent components. We used the algorithm to estimate C independent components, in which C is the number of dimensions explaining 99.5% of the variance of a dataset.

To estimate statistically small sets of chromatophores receiving common drive ('motor elements', see main text), we subsequently clustered the small subset of chromatophores with high values on single independent components of the matrix S. Because motor-unit membership is binary (a chromatophore either is innervated by a motor neuron or is not), we thresholded the independent components to extract these highly contributing chromatophores and examine their properties. We found that the highest contributing chromatophores most often clustered spatially in single modes, with chromatophores contributing less located further away. We chose our threshold for motor-element inclusion such that the median of the chromatophore spatial distribution matched approximately that measured in vitro. We assigned a sign to each independent component as the sign of the maximum value (chromatophore) of that independent component. Values higher than 8 s.d. above the mean value of the positive independent components or lower than 8 s.d. below the mean value of the negative independent components were clustered to form a motor element. We then visually inspected the motor elements to check for colour classification errors and to remove groups that did not contain well-segmented chromatophores due to errors in watershedding.

Inference of putative motor control hierarchy. We first averaged the filtered, downsampled area time series (as in 'Inference of motor elements from in vivo imaging data') for all chromatophores assigned to a motor element (ignoring the sign or weight of its associated independent component). This procedure was motivated by the known underlying biology: it attempted to approximate the common motor neuron drive that caused the chromatophores to be clustered into a motor element. Note that the precision of this approximation depends on several factors, including the multiple and partially overlapping innervation of chromatophores, the difficulties of inferring motor units (as described above) and the fact that the relationship between chromatophore size and neural drive is likely to be sigmoid, and thus linear only in a limited range. We performed agglomerative hierarchical clustering of these time series of motor elements, using the correlation coefficient as a distance metric and complete linkage. To segment monochromatic clusters at different levels of the hierarchy, we measured the fraction of clusters composed of motor elements that contained only light or only dark chromatophores.

Chromatophore colour changes over development. For precise characterization of chromatophore colours (Fig. 5a, b), we took images of two cuttlefish over days using an 18-M-pixel camera (Canon, 550D) at 10–18× magnification. Recognizable landmarks (for example, papillae and mantle edges) were used to return to the same area of skin repeatedly. We aligned skin patches using TrakEM2 (ImageJ plugin). Images of chromatophores were white-balanced using a nearby patch of skin that did not contain any chromatophore as a reference. The colour of individual chromatophores was then determined in hand-drawn regions of interest (ROIs). To visualize the colour change in a condensed representation in colour space, the three colour channels (red, green and blue) were averaged over all pixels within the ROI. Colours were then converted from RGB into hue-saturation-value colour space, which assigns brightness and hue to different axes. To determine an average trajectory in colour space, the chromatophores were temporally aligned to the transition state by thresholding on the red colour channel.

Chromatophore sizes over development. To check for potential size and variability differences of the transition state, we aligned a dataset separated by two days using our pipeline and identified transitioning chromatophores as those that were classified as light on the earlier day and dark on the later one. Yellow chromatophores were defined as light chromatophores that were not transitioning. Size and variability were estimated using filtered data (as in 'Inference of motor elements from in vivo imaging data'). For validation independent of our tracking pipeline (Extended Data Fig. 9c), we aligned images using TrakEM2, and manually grouped individual chromatophores into three categories (yellow/orange, reddish-brown, black) without temporal context. We then incorporated developmental information by retaining only those yellow/orange chromatophores that were observed again at a later time point as yellow/orange, that is, those data, that were not close to the transition. Similarly, black chromatophores were retained only if they were observed earlier already as black. Size was determined on hand-drawn ROIs outlining each chromatophore in the aligned dataset.

Colour–development–numbers model. For the juvenile animal, the generation rate of new chromatophores was estimated by counting, in datasets aligned over days, all chromatophores within a patch of skin on the last day. We then found the fraction of these chromatophores that was present on previous days within the same aligned patch. The birth rate was calculated as an exponential fit to these data, using tenfold cross-validation of skin regions. The estimated ratio of light/dark was counted from manual annotation of a patches of skin taken at high-resolution ($n=7$ animals). In the adult animal, light-to-dark chromatophore transition took longer than our 42-day observation period. The derivation of our model and the method used to estimate light-chromatophore lifetime are provided in full in the Supplementary Information.

Growth model. The model described in Fig. 6 is illustrated in more detail in Extended Data Fig. 9a. For simplicity, growth of a skin patch was modelled by a sequence of three steps, repeated every 'day': (1) insertion of new chromatophores; (2) isotropic expansion of skin patch; and (3) age-dependent size change of inhibitory surrounds and updating of the chromatophore-colour label. First, we explain how the inhibitory surround was constructed from observations. Second, we define the relevant parameters. Finally, we describe the simulation steps in detail and explain how parameters affect simulation outcome.

The zone of inhibition surrounding individual chromatophores in our spatial model was generated from the empirical average of chromatophore density surrounding a chromatophore ($n = 9{,}199$ chromatophores of both colours, $n = 1$ animal). We normalized the radial average of the density by the value of the first peak and set any value occurring at greater radial distance to 1. We then fitted a logistic function $I(x) = \frac{1}{1+e^{-k(x-s_0)}}$ to this density, in which x is the distance from the chromatophore centre, s_0 the size (that is, distance at half-height) and k the slope at half-maximum. We inverted this function as $1 - I(x)$, to arrive at the one-dimensional inhibitory surround kernel (Extended Data Fig. 9d). This curve defined the radial dependence of the isotropic two-dimensional inhibitory surround $J(\boldsymbol{x}) = \frac{1}{1+e^{k(\boldsymbol{x}-s_0)}}$, with vector $\boldsymbol{x} = (x_1, x_2)$ denoting the two-dimensional spatial coordinates.

With the shape of the surround fixed, the simulation contained five parameters: (1) the maturation age of chromatophores (L–D transition day); (2) the rate at which chromatophores move away from each other daily (skin growth rate); (3) the size of the inhibitory surround of a chromatophore at birth (s_b); (4) the rate of change of the inhibitory surround as a chromatophore ages (r_a); and (5) the threshold level of skin 'filling' at which a 'day' is complete. Note that these five parameters can all be varied independently of each other. Extended Data Figure 9a shows the simulation steps at which each parameter is introduced.

Our analytical growth model (above, and Supplementary Information) determines the coupling of maturation age (parameter 1), the rate of new chromatophore insertion (approximately parameter 2 squared) and the L/D ratio. We therefore fixed the values for parameter 1 and 2 approximately to values observed in a young animal and, as expected, observed a realistic L/D ratio. Parameters 3–5 determined the local spatial layout of chromatophores. Parameter 5 allowed us to model the possibility that chromatophores are not inserted to maximum packing density on any given day. Extended Data Figure 9 illustrates that this parameter affects chromatophore packing without affecting L–D interdigitation. Our results did not depend on the number of simulated 'days', provided that simulations were run for long enough (around 60 'days') to allow L/D to reach steady state.

We defined an 'inhibition' field $F(\boldsymbol{x})$ over the simulated patch of skin of initial size $l \times l$ ($l = 540$ μm), in which both components of $\boldsymbol{x} = (x_1, x_2)$ are real numbers in the interval $\left[-\frac{l}{2}, \frac{l}{2}\right]$. The inhibition field was computed as $F(\boldsymbol{x}) = \sum_i J_i(\boldsymbol{x} - \boldsymbol{x}_0^i)$, that is, by adding the inhibitory kernels J_i (see above) from all present chromatophores i with position $\boldsymbol{x}_0^i$ and size s_0^i (determined by parameters 3 and 4). Values larger than 1 were set equal to 1. F was updated after every chromatophore insertion. The simulation began with an empty patch of skin, that is, with an inhibition field equal to 0 (uniform probability for chromatophore insertion). During a simulation 'day', chromatophores were inserted sequentially. A location was drawn from a two-dimensional uniform random distribution over the space covered by the skin patch, and insertion took place with probability $P = 1 - F$ at that location. The 'day' ended when no location in the inhibition field was left with a value less than the filling threshold (parameter 5). At the end of that 'day', the system was expanded by scaling all positions $\boldsymbol{x}$ by a fixed rate (parameter 2). The size of the inhibitory surround was then adjusted according to chromatophore age (a) as $s_0 = s_b(1 + r_a a)$, (until $a = 19$ days, after which s_0 was fixed), and the colour of each chromatophore was updated.

We fitted parameters 3–5 to the radially averaged chromatophore-triggered densities from one animal ($n = 4{,}095$ dark, 5,104 light) using a grid search and mean-squared-error loss. Search space was: (3) 67–81 μm radius (full width, half maximum); (4) −3.2% to −1.6% per 'day' (adjusted every two 'days'); and (5) 0–0.5 filling threshold.

The parameters of the best fit model were (1) 19 days; (2) 2.06% per 'day'; (3) 75.6-μm radius (full width, half maximum); (4) −2.8% per 'day' (adjusted every two 'days'); (5) 0.1. To approximate squid skin (Extended Data Fig. 10), we adjusted parameter 4 to 2.8% per 'day' (adjusted every two 'days') until 45 days of age followed by an expansion to 340 μm radius.

Image manipulation. Colour images in Figs. 1c, d, 3c, 5e were uniformly and linearly scaled for clarity.

Statistics. Unless stated otherwise, data are mean ± s.d. For box plots, central line indicates the median; box limits are quartiles. Whiskers extend to a maximum of ±2.7s.d. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

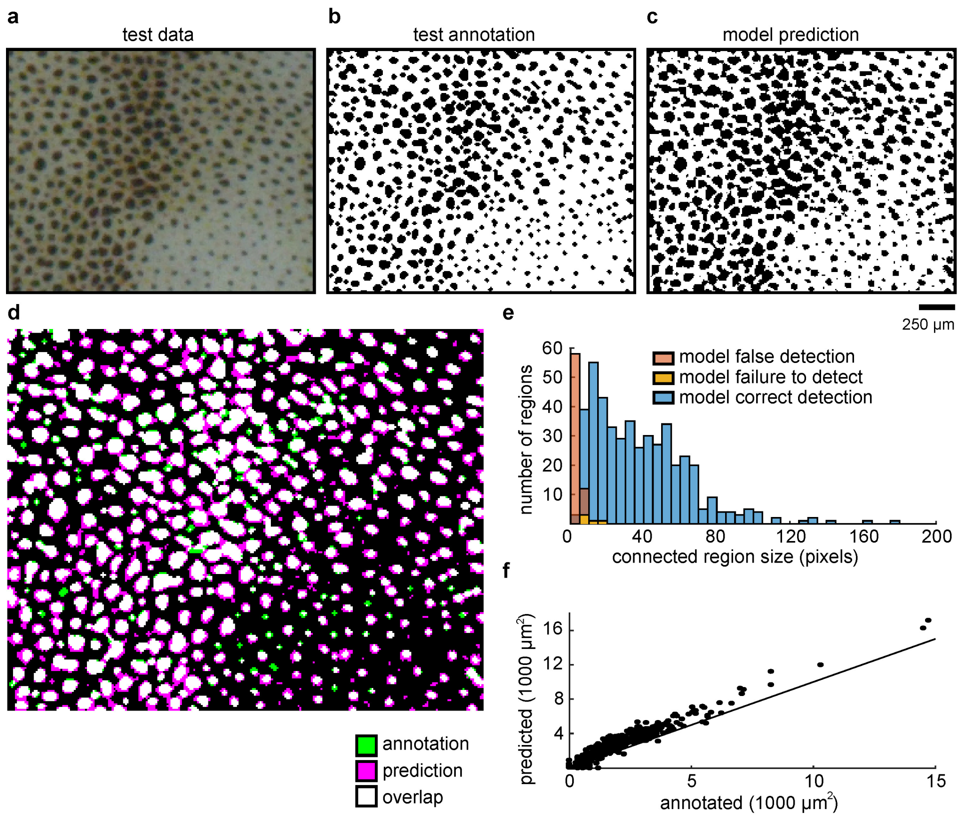
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The code developed in this study is posted in a repository on GitHub: https://github.molgen.mpg.de/MPIBR/cuttlefish-code-nature.

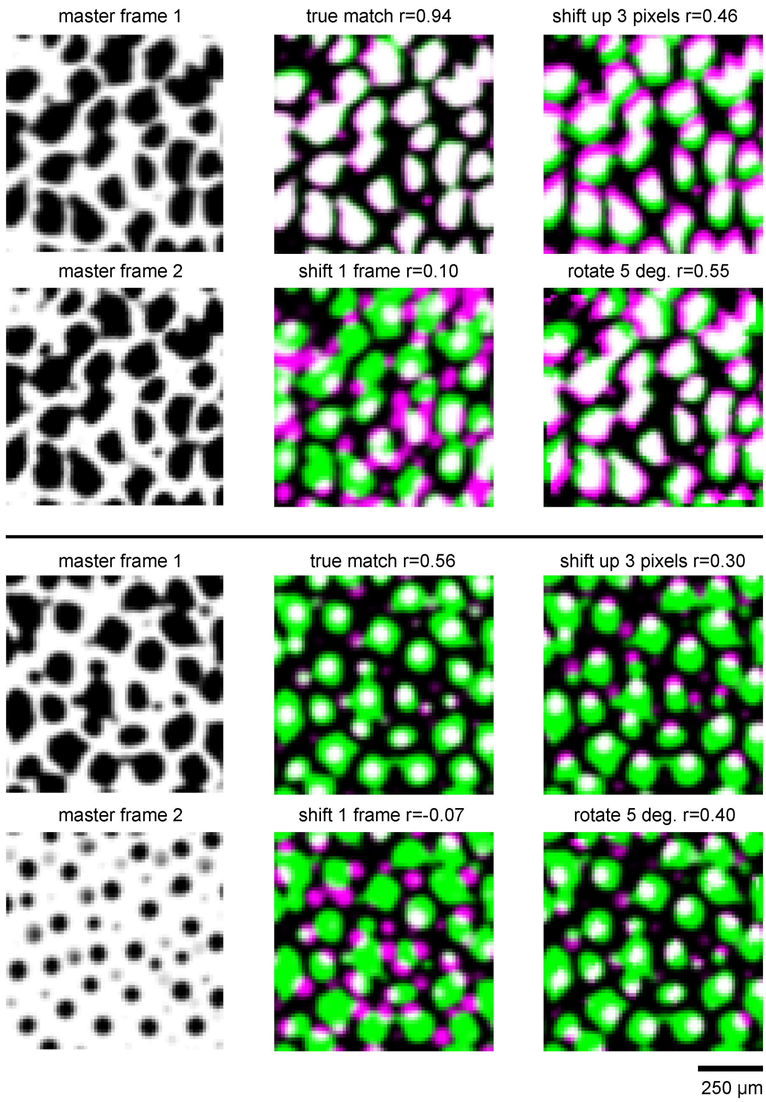
Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

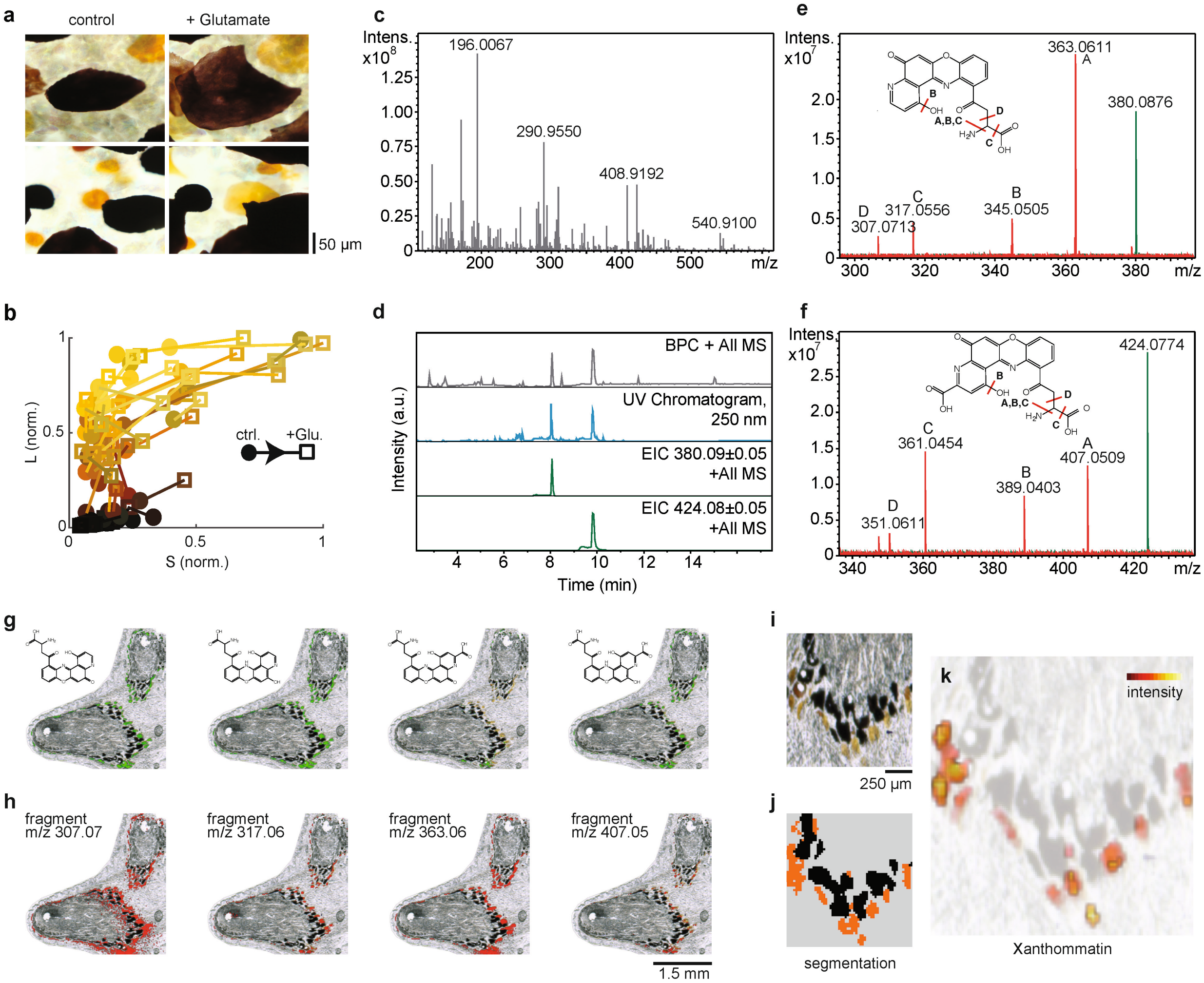
42. Butler-Struben, H. M., Brophy, S. M., Johnson, N. A. & Crook, R. J. In vivo recording of neural and behavioral correlates of anesthesia induction, reversal, and euthanasia in cephalopod molluscs. *Front. Physiol.* **9**, 109 (2018).
43. Bergstra, J., Yamins, D. & Cox, D. D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proc. 30th International Conference on Machine Learning* Vol. 28 I-115–I-123 (Journal of Machine Learning Research, 2013).
44. Schaefer, S., McPhail, T. & Warren, J. Image deformation using moving least squares. *ACM Trans. Graph.* **25**, 533–540 (2006).
45. Fischler, M. A. & Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**, 381–395 (1981).
46. Meyer, F. Topographic distance and watershed lines. *Signal Process.* **38**, 113–125 (1994).
47. Yoo, A. B., Jette, M. A. & Grondona, M. in Job Scheduling Strategies for Parallel Processing Vol. 2862 (eds Feitelson, D. et al.) 44–60 (Springer, Berlin, 2003).
48. Message Passing Interface Forum. MPI: a message-passing interface standard. version 3.1 http://mpi-forum.org/mpi-31/ (2015).
49. Jones, E. et al. SciPy: open source scientific tools for Python. http://www.scipy.org/ (2001).
50. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
51. Walt, S. et al. scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).



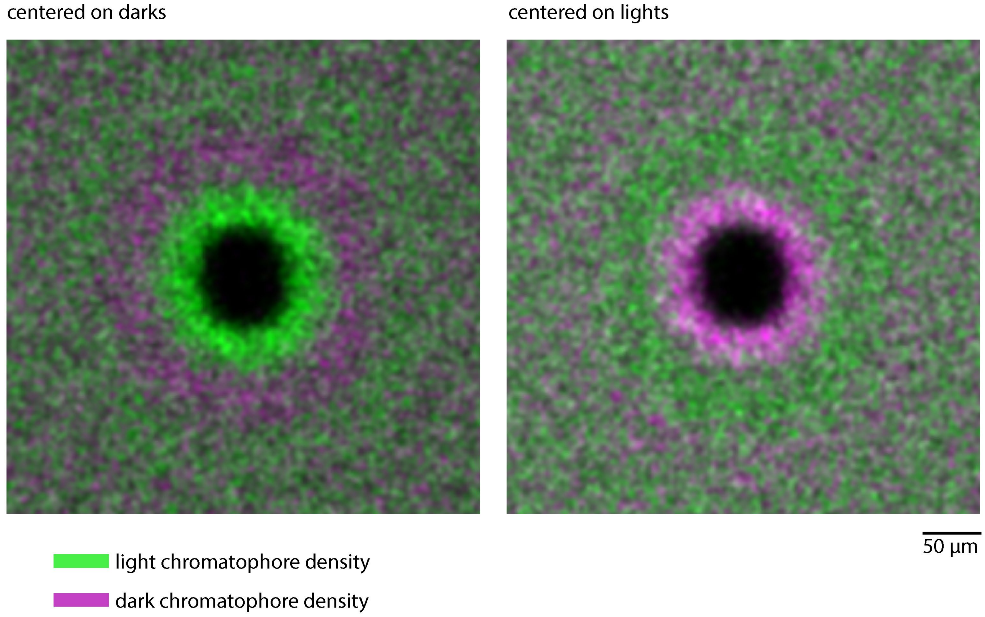
Extended Data Fig. 1 | Accuracy of the chromatophore classifier. **a**, Test patch of skin used for classifier testing. **b**, Segmentation by expert human. **c**, Segmentation by classification algorithm. **d**, Composite image comparing manual (annotation) and automatic (prediction) segmentation. There was agreement for 87% of pixels, with differences mostly on the edges surrounding chromatophores. **e**, Quantification of region overlap. Regions defined from watershedding the composite image shown in **d**. Correct detection: regions labelled by both methods. False detections: regions identified by automatic but not by manual segmentation. Failed detections: regions identified by manual but not automatic segmentation. **f**, For all regions, annotated versus predicted size. Line, identity.



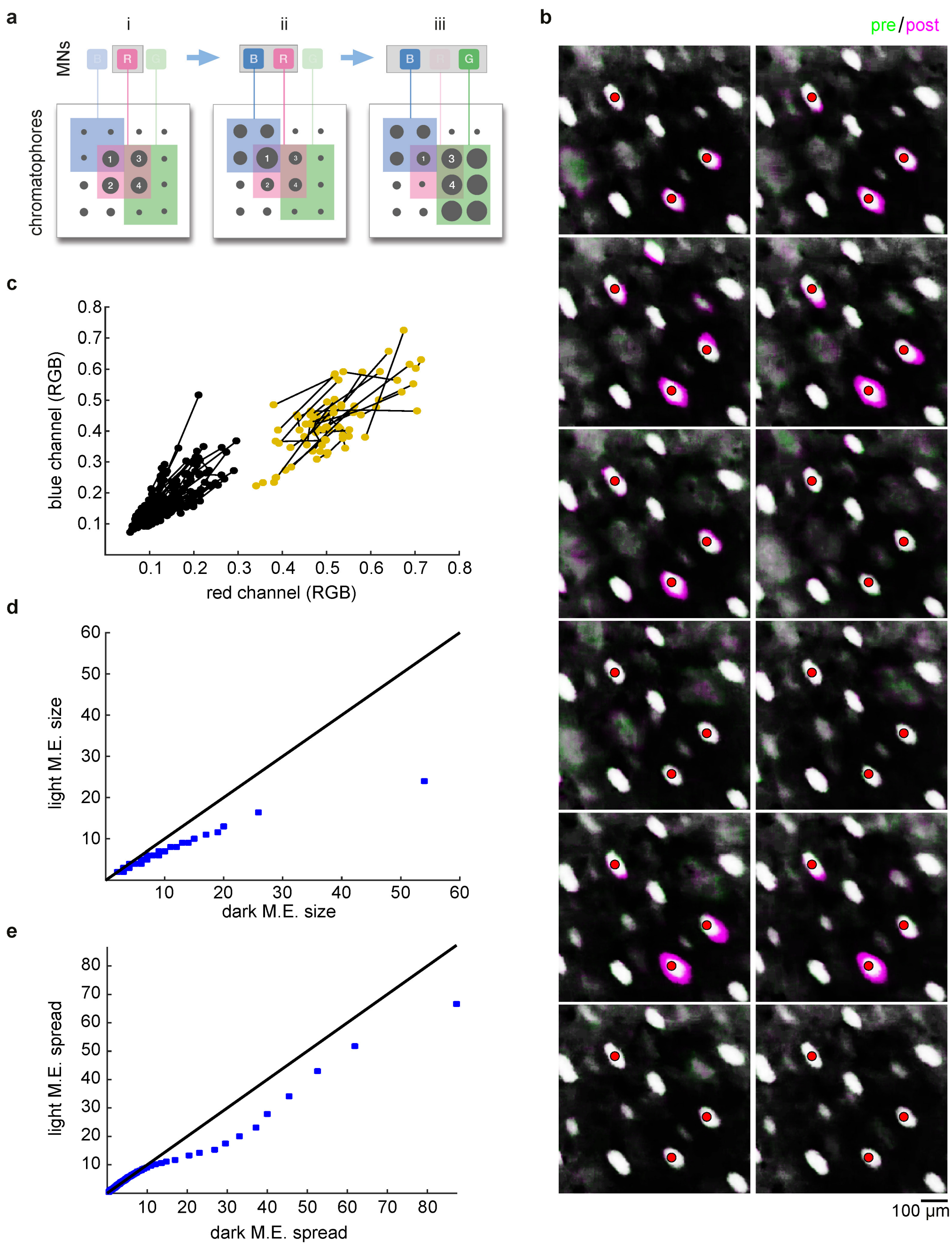
Extended Data Fig. 2 | Sensitivity of the correlation between skin patches to small image translations and rotations. Left, skin patches from the two sets of matching master frames shown in Fig. 1e. Middle and right, composite images of the corresponding master frames (master frame 1 in green, master frame 2 in magenta, overlap in white). Small translations or rotations quickly lower the cross-correlation, as in the schematic in Fig. 1e, bottom.



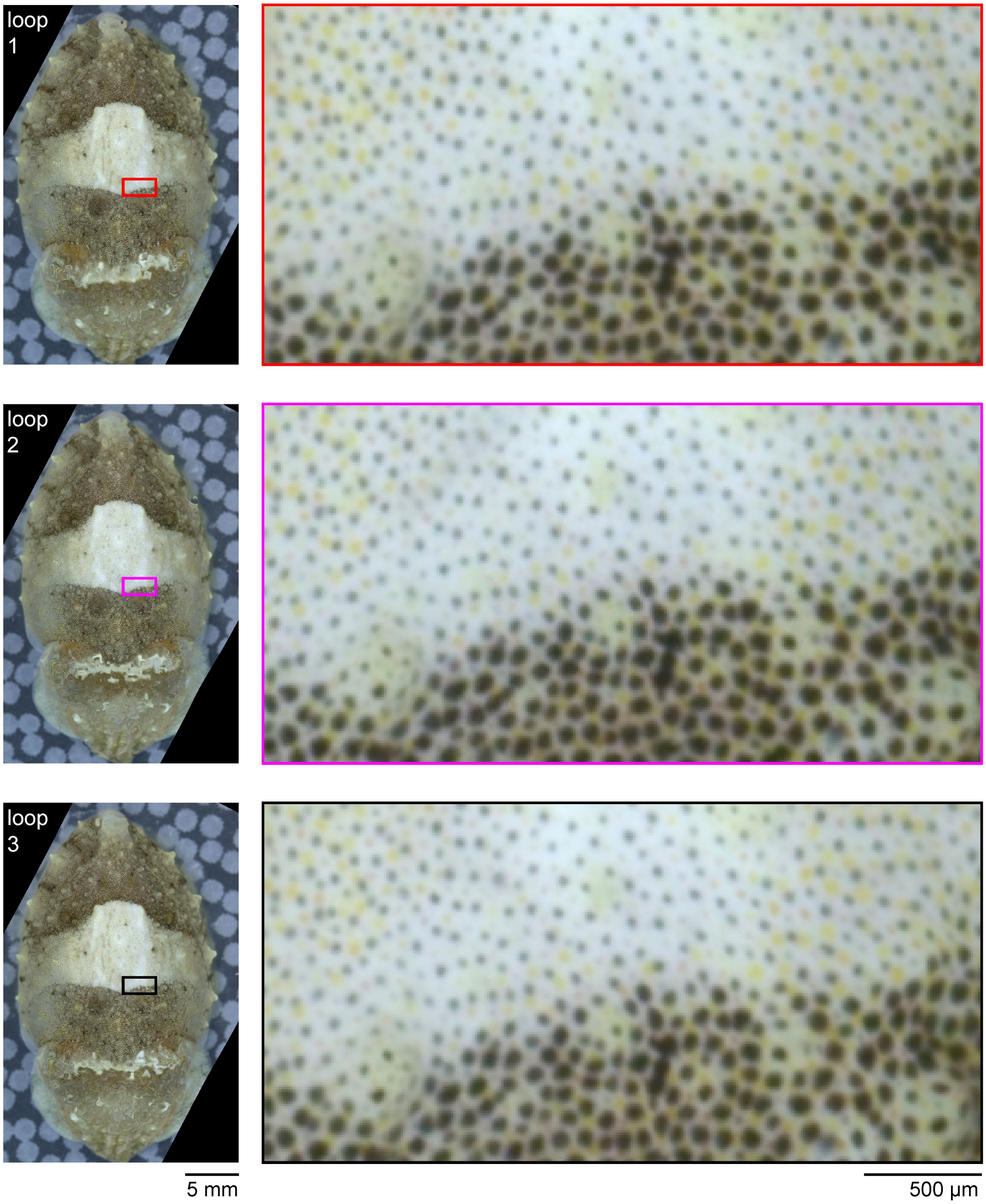
Extended Data Fig. 3 | Identification and localization of xanthommatin in light chromatophores. **a**, Chromatophores before (left) and after (right) local application of 40 µM glutamate. **b**, Transmission spectra of a population of chromatophores before (circles) and after (squares) glutamate application, projected onto two dimensions defined by human L and S cone action spectra ($n = 63$ chromatophores). **c**, Direct infusion electrospray ionization Fourier-transform ion cyclotron resonance (ESI-FT-ICR) mass spectrum of the skin tissue extract showing high spectral complexity. **d**, HPLC–UV–MS chromatograms of skin tissue extract showing two main peaks with correlating ultraviolet-light (250 nm) absorption (blue) and mass spectrometry intensity (grey) consisting of eluting compounds with *m*/*z* 380.09 and *m*/*z* 424.08 (extracted ion chromatogram (EIC) traces, green). Experiments were replicated five times with similar results. **e**, Direct infusion ESI-FT-ICR mass spectra of skin tissue extract showing an overlay of the isolated precursor spectrum for decarboxylated xanthommatin (green, *m*/*z* 380.0876, theoretical: *m*/*z* 380.0877) and the fragment spectrum (red). Main fragments were assigned to putative structural losses of A (-NH_3), B (-H_2O, -NH_3), C (-NH_3, -HCOOH), D (-$C_2H_3NO_2$) by accurate mass. **f**, Direct infusion ESI-FT-ICR mass spectra of skin tissue extract showing an overlay of the isolated precursor spectrum for xanthommatin (green, *m*/*z* 424.0774, theoretical: *m*/*z* 424.0775) and the fragment spectrum (red). Main fragments were assigned to putative structural losses of A (-NH_3), B (-H_2O, -NH_3), C (-NH_3, -HCOOH), D (-$C_2H_3NO_2$) by accurate mass. **g**, Intensity distributions in laser desorption ionization Fourier-transform ion cyclotron resonance mass spectrometry (LDI-FT-ICR-MS) imaging and structures for putative xanthommatin derivatives (merged $[M + H]^+$; $[M + Na]^+$): decarboxylated, oxidized (*m*/*z* 380.0886; 402.0696), decarboxylated, reduced (*m*/*z* 382.1037; 404.0853), oxidized (*m*/*z* 424.0785; 446.0629) and reduced (*m*/*z* 426.0938; 448.0761). **h**, Intensity distributions of main xanthommatin and derivative fragments, corresponding to molecular species detected in ESI-FT-ICR fragmentation measurements. Experiments were performed on 12 tissue slices, producing similar results. **i**, Image of cryotome section of fresh-frozen *Sepia* skin showing chromatophores. **j**, Spatial segmentation map of section in **i**, showing distinct clusters for light and dark chromatophores (orange versus black colours) and surrounding tissue (grey). **k**, Intensity distributions for xanthommatin derivatives (merged $[M + H]^+$ and $[M + Na]^+$) obtained from LDI-FT-ICR-MS imaging experiments ($n = 1$ sample).



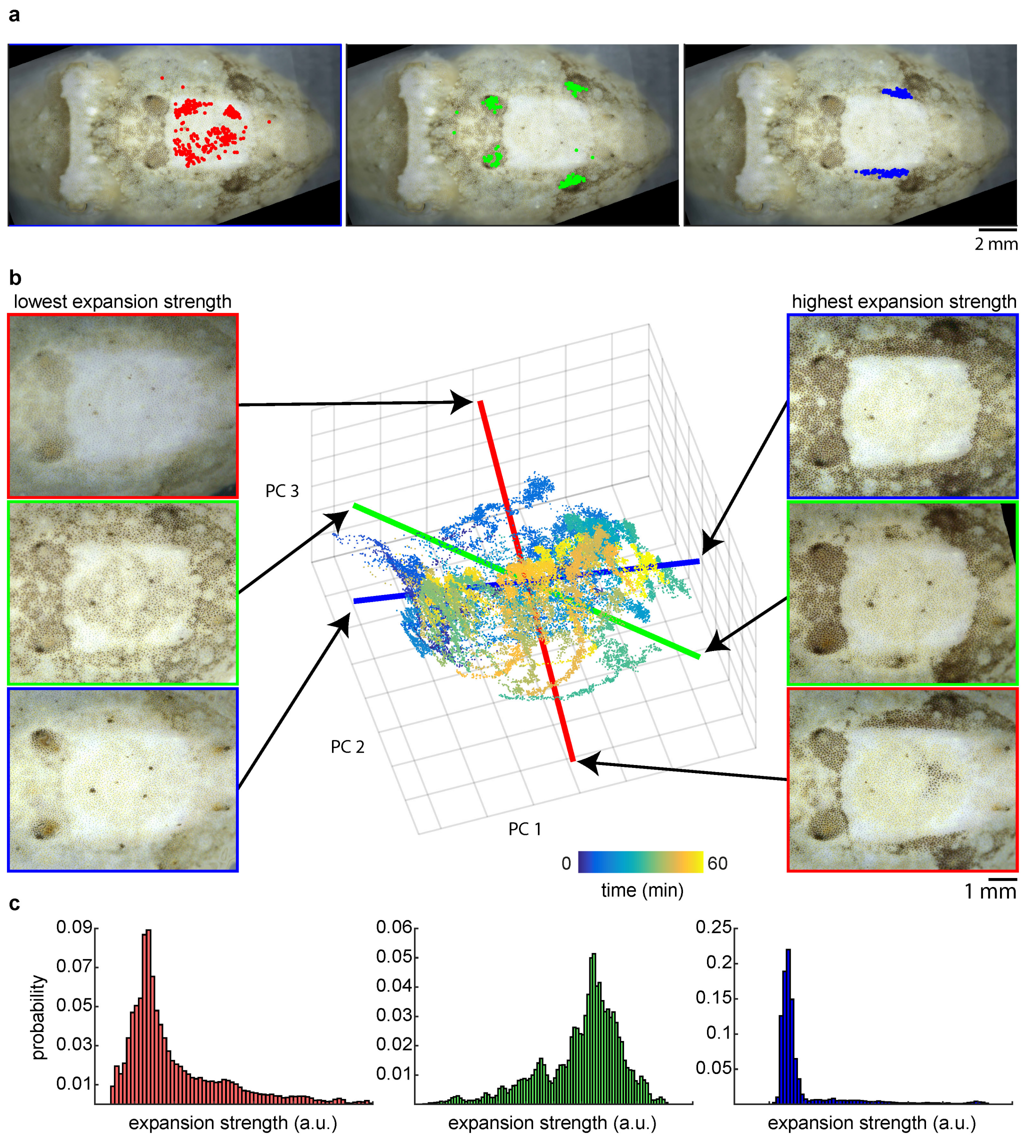
Extended Data Fig. 4 | Chromatophore-centred average densities. Two-dimensional density distributions for light and dark chromatophores over the mantle of an animal ($n = 9,199$ chromatophores). The composite images show the density of light chromatophores in green and the density of dark chromatophore in magenta. For visualization, densities were linearly scaled together within an image. This preserves relative densities within each image but leads to slightly different colours across images.



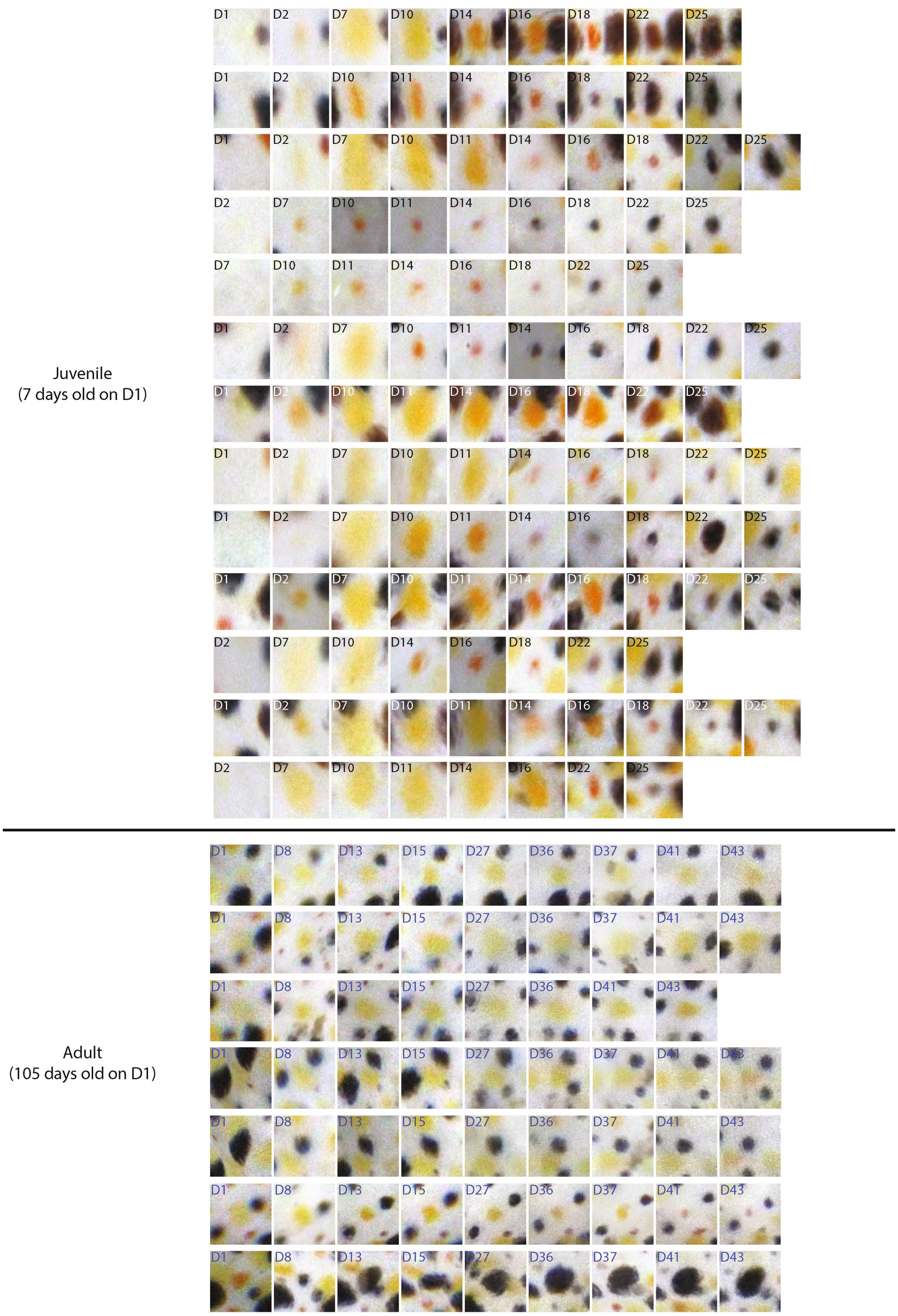
Extended Data Fig. 5 | Identification of motor units. a, Schematic showing three hypothetical, partially overlapping motor units (defined by motor neurons (MNs) B, R and G), tracked over three epochs (i–iii), each characterized by different co-activation patterns (epoch i, R alone; ii, B + R; and iii, B + G). Even though chromatophores 1–4 all belong to the same motor unit (R), their average pairwise correlation during these three epochs would differ owing to the activity of the partially overlapping motor units B and G; identifying motor units using this metric would thus fail. This hypothetical example indicates that the units of coordination during behaviour could be smaller than single anatomical motor units (they could also be larger; for example, if some motor neurons are always centrally coupled). **b**, Single trials of minimal electrical stimulation experiments in in situ nerves. Composite images (one per trial), green, 10 ms pre-stimulus; magenta, 200 ms post-stimulus; white, overlap. Threshold stimulation either leads to the expansion of a set of three chromatophores (marked with red circles, for example, trial 1), or fails to activate any chromatophore (for example, trial 6, 114 motor units determined with this method). **c**, Colour assignment of chromatophores in situ. Colour label was assigned based on a threshold on the red channel of RGB space (0.3). Chromatophores (dots) belonging to the same motor unit (as determined in **a**) are connected by lines, revealing the monochromaticity of motor units. $n = 114$ chromatophores. **d**, Dark motor elements tend to be larger than light motor elements. *Q*–*Q* plot showing quantiles of the dark versus light motor element size distribution. Line, identity. **e**, Tail of distribution of motor element spread is heavier with dark than light chromatophores. *Q*–*Q* plot showing quantiles of the dark versus light motor element spread (calculated as in Fig. 3d). Line, identity.



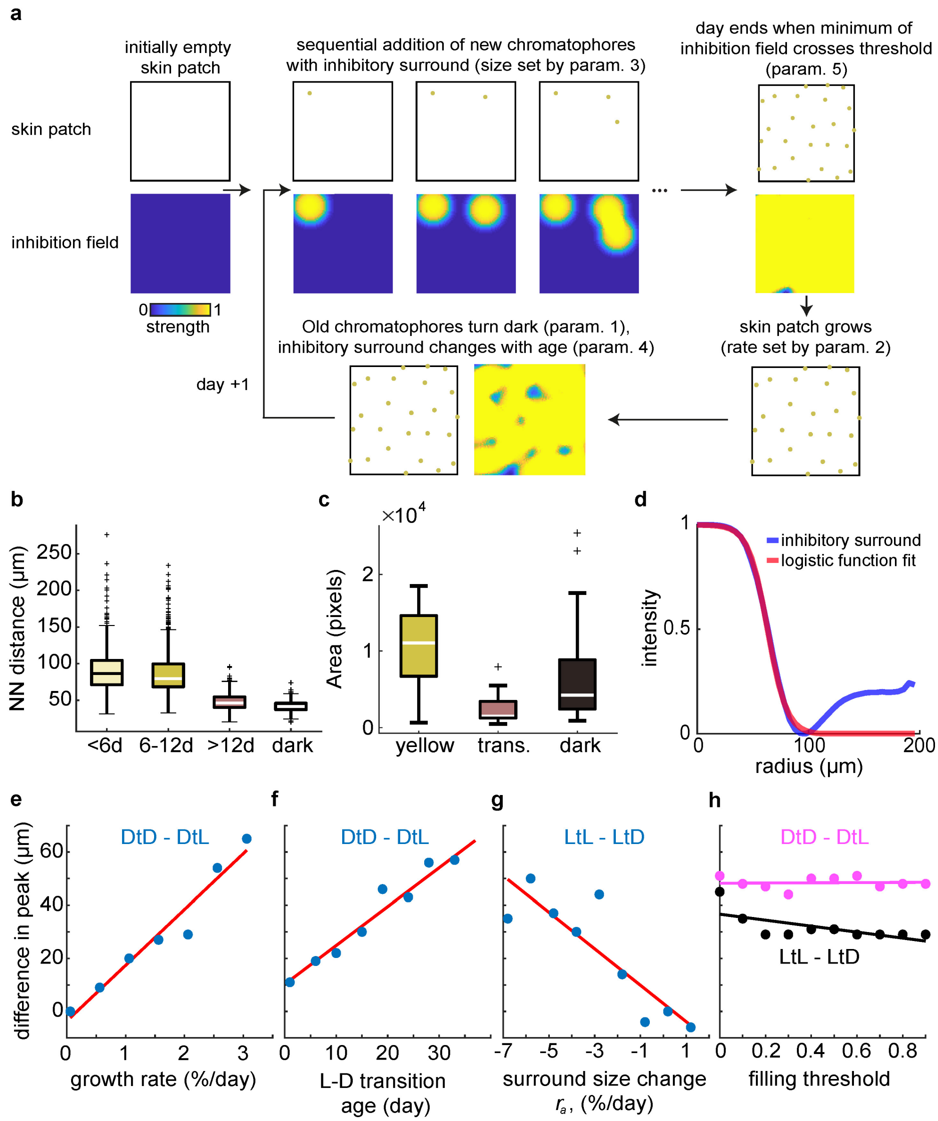
Extended Data Fig. 6 | Pattern-border precision at single-chromatophore level. Left, three similar points along the pattern trajectories shown in Fig. 4b after chromatophore alignment. Right, expanded view of a pattern border. Note the remarkably similar expansion states of the chromatophores at each of the three visits, and the rugged pattern borders at chromatophore scale, with interdigitation of expanded and contracted chromatophores, generating apparent noise. This apparent noise may be critical for natural realism.



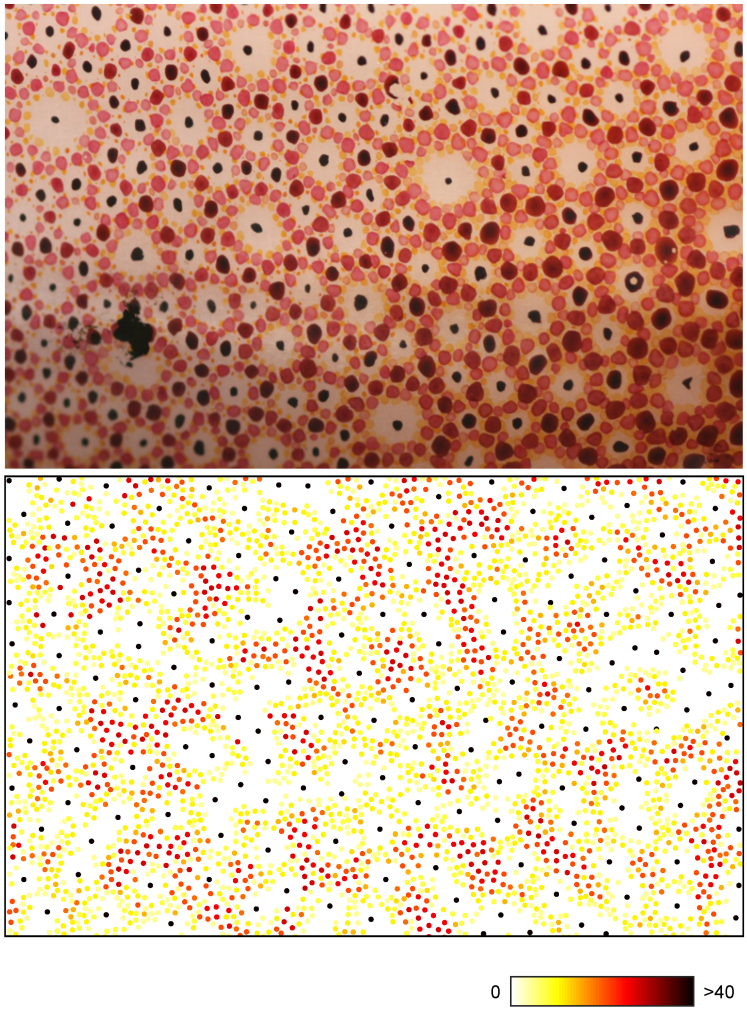
Extended Data Fig. 7 | Linking statistical hierarchy of pattern elements to dynamics. a, Three example intermediate-level clusters of motor elements (threshold of 0.4 as in Fig. 3i, different animal), overlaid on the average aligned colour image for the dataset (216,160 images). The clusters are mostly composed of chromatophores of a single colour: cluster 1 (red) is light; clusters 2 and 3 (green and blue) are dark. **b**, The dynamics of a 60-min dataset, projected onto the first three principal components (48% variance explained, $n = 1{,}437$ chromatophores, 52,040 samples). A cluster activity direction can be defined in principal component space by projecting the cluster identity vector (vector of length = number of chromatophores, with 1s assigned to chromatophores in a cluster, otherwise it is assigned 0), onto the principal components. The coloured lines show the cluster activity directions for the three clusters in **a**. Projecting the dataset onto these directions shows the expansion strength of the cluster at different times. The images corresponding to the times of lowest and highest strengths are shown to the left and right, respectively. **c**, Full distribution of expansion strengths, projecting all time points onto cluster activity directions. In this dataset, cluster 2 is often expanded, whereas clusters 1 and 3 are rarely expanded. a.u., arbitrary units.



Extended Data Fig. 8 | Chromatophores change colour from light to dark as they age. A gallery of aligned patches of skin centred on the position of chromatophore insertion is shown. Top, juvenile animal, 7 days old on the first day of observation (D1). Left-most column shows skin pre-chromatophore-birth. Over approximately 19 days, chromatophores that first appear pale yellow darken progressively, transitioning to orange and red, before finally turning black. Field of views (FOVs): from around 150×150 µm on day 1 of observation to 300×300 µm on day 25 of observation. Bottom, adult animal, 105 days old on day 1 of observation. Rows show chromatophores undergoing a similar light–dark colour transition as in the juvenile (top), but at a much slower rate. FOVs: around 200×200 µm (nonlinear alignment). Examples were chosen from aligned skin patches containing around 100 chromatophores.



Extended Data Fig. 9 | Development of the chromatophore array. **a**, Flowchart depicting the spatial-growth-model algorithm and highlighting the involvement of model parameters (Methods). **b**, Box plots of nearest-neighbour (NN) distances between young (<6 days old) and older chromatophores. Young chromatophores are significantly closer to both older light (>12 days) and dark chromatophores than to other young or middle-aged (6–12 days) light chromatophores. ($P < 0.0001$, Kruskal–Wallis followed by Tukey's HSD, $n_{\text{chromatophores}} = 522$ for <6 days, 541 for 6–12 days, 1,550 for >12 days, 1,910 dark chromatophores, 1 animal). Distances calculated on a single image, ages estimated by finding the day of chromatophore birth on aligned developmental datasets (Methods). **c**, Distributions of size for yellow, red (transitional (trans.)) and dark chromatophores, annotated manually (validation of analysis in Fig. 5d). Transitional chromatophores are significantly smaller than either yellow or dark ones (transitional versus yellow, $P = 1.0 \times 10^{-7}$; transitional versus dark, $P = 6.3 \times 10^{-4}$; $n = 70$ yellow, 16 transitional, 84 dark chromatophores; two-tailed Wilcoxon rank-sum tests, $n = 1$ animal). Box plots show the central line, median; box limits, quartiles; whiskers, ±2.7 s.d. **d**, Generation of the inhibitory surround used in the skin growth model (Fig. 6b). Blue, empirical radially averaged chromatophore centred density, inverted and normalized 0:1. Red, logistic function fit to the blue density, as in Fig. 6b. **e**–**h**, Manipulating single parameters of the skin growth model suggests the mechanisms underlying colour interdigitation. **e**, Difference between peak dark-triggered dark-chromatophore density (DtD) and dark-triggered light-chromatophore density (DtL), as a function of model skin growth rate. Points are from the average of three model runs. Line, linear fit. ANOVA F-statistic versus the constant model $F = 96.6$, $P = 0.000186$. **f**, Difference between peaks of radially averaged dark-triggered dark-chromatophore density and dark-triggered light-chromatophore density, as a function of age at which chromatophores transition from light to dark. Points are from a single model run, in which the colour class was changed according to chromatophore age. Line, linear fit. F-test for linear regression: $F = 152$, $P = 5.26 \times 10^{-6}$. **g**, Difference between first peak (first zero-crossing of derivative of radially averaged density) in the radially averaged light-triggered light-chromatophore density (LtL) and light-triggered dark-chromatophore density (LtD), as a function of r_a, the rate at which the inhibitory surround changes with chromatophore age. Points are from the average of three runs of the model. Line, linear fit. F-test for linear regression: $F = 21.9$, $P = 0.00226$. **h**, Colour interdigitation is robust to stop-criterion used to define end of 'day' (parameter 5, Methods). Magenta, DtD − DtL (as in **e**, **f**). Black, LtL − LtD. Lines, linear fits. F-test for linear regression: $F = 0.0206$, $P = 0.889$ (DtD − DtL); $F = 6.57$, $P = 0.0334$ (LtL − LtD). Points in **e**–**h** are from the average of three model runs.



Extended Data Fig. 10 | Exploration of developmental-model parameters reveals species-specific patterns. Changing model parameters (see main text and Methods) can lead to the characteristic rings observed in some squid species, with single light chromatophores at the centre and a radial centrifugal darkening gradient. Top, skin of common squid, *Loligo vulgaris* (image by R. Siegel). Bottom, simulation of development using a profile of change of the inhibitory disc centred on each chromatophore r_a different from that used in Fig. 6b for *S. officinalis*.

https://doi.org/10.1038/s41586-018-0590-4

Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*

The Tabula Muris Consortium*

Here we present a compendium of single-cell transcriptomic data from the model organism *Mus musculus* that comprises more than 100,000 cells from 20 organs and tissues. These data represent a new resource for cell biology, reveal gene expression in poorly characterized cell populations and enable the direct and controlled comparison of gene expression in cell types that are shared between tissues, such as T lymphocytes and endothelial cells from different anatomical locations. Two distinct technical approaches were used for most organs: one approach, microfluidic droplet-based 3′-end counting, enabled the survey of thousands of cells at relatively low coverage, whereas the other, full-length transcript analysis based on fluorescence-activated cell sorting, enabled the characterization of cell types with high sensitivity and coverage. The cumulative data provide the foundation for an atlas of transcriptomic cell biology.

The cell is a fundamental unit of structure and function in biology, and multicellular organisms have evolved various cell types with specialized roles. Although cell types have historically been characterized by morphology and phenotype, the development of molecular methods has enabled increasingly precise descriptions of their properties, typically by measuring protein or mRNA expression patterns[1]. Technological advances have also expanded measurement multiplexing such that highly parallel sequencing can now enumerate nearly every mRNA molecule in a single cell[2–8]. This approach has provided insights into cell biology and organ composition from various organisms[9–18]. However, although these reports provide valuable characterization of individual organs, it is challenging to compare data collected from different animals by independent labs with varying experimental techniques. It therefore remains unknown whether these data can be synthesized as a more general resource for biology.

Here we report a compendium of cell types from the mouse *Mus musculus*; we refer to this as a *Tabula Muris*, or 'Mouse Atlas'. We analysed several organs from the same mouse, generating a dataset controlled for age, environment and epigenetic effects. This enabled the direct comparison of cell-type composition between organs, and the comparison of shared cell types across organs. The compendium comprises single-cell transcriptomic data from 100,605 cells isolated from 20 organs from three female and four male, C57BL/6JN, three-month-old mice (10–15 weeks), analogous to 20-year-old humans (Fig. 1a). Aorta, bladder, bone marrow, brain (cerebellum, cortex, hippocampus and striatum), diaphragm, fat (brown, gonadal, mesenteric and subcutaneous), heart, kidney, large intestine, limb muscle, liver, lung, mammary gland, pancreas, skin, spleen, thymus, tongue and trachea from the same mouse were immediately processed into single-cell suspensions. All organs were single-cell-sorted into plates using fluorescence-activated cell sorting (FACS), and many were also loaded into microfluidic droplets (see Extended Data and Methods).

All data, protocols, analysis scripts and an interactive data browser are publicly available (for details, see 'Data availability'). This release enables the exact replication of all results, facilitates in-depth analyses not completed here, and provides a comparative framework for future studies using the large variety of murine disease models. Although these data are by no means a complete representation of all mouse organs and cell types, they provide a first draft attempt to create an organism-wide representation of cellular diversity.

Defining organ-specific cell types

To define cell types, we analysed each organ independently by performing principal component analysis (PCA) on the most variable genes between cells, followed by nearest-neighbour graph-based clustering. We then used cluster-specific gene expression of known markers and genes that are differentially expressed between clusters to assign cell-type annotations to each cluster (Extended Data Figs. 1, 2, Supplementary Table 1). We used a standard annotation method for all organs; step-by-step instructions to reproduce this method are provided in the supplemental Organ Annotation Vignette using the liver as an example. Cell type descriptions and defining genes for each organ are available in the Supplementary Information. For each cluster, we provide annotations in the controlled vocabulary of the cell ontology[19] to facilitate inter-experiment comparisons. Many of these cell types have not previously been obtained in pure populations, and our data provide a wealth of new information on their characteristic gene-expression profiles. Some unexpected discoveries include a potential new role for *Neurog3*, *Hhex* and *Prss53* in the adult pancreas, a cell population expressing *Chodl* in limb muscle, transcriptional heterogeneity of brain endothelial cells, the expression of MHC class II genes by adult mouse T cells, and sets of transcription factors that distinguish cell types across organs.

Methodological comparison

We performed single-cell RNA-sequencing with two methods: FACS-based cell capture in plates and microfluidic-droplet-based capture (hereafter denoted the FACS method and the microfluidic-droplet method, respectively). To understand the technical biases of each approach, we performed both methods on many organs. Overall, 44,949 cells from the FACS method and 55,656 cells from the microfluidic-droplet method were retained after quality control. Single-cell transcriptomes were sequenced to an average depth of 814,488 reads per cell (FACS) and 7,709 unique molecular identifiers (UMIs) per cell (microfluidic droplet). Comparing methods shows organ-specific differences in the number of cells analysed (Fig. 1b, c), reads per cell (Extended Data Fig. 3a, c) and genes per cell (Extended Data Fig. 3b, d). Furthermore, with both methods the most abundant cell types analysed were epithelial cells and leukocytes, although FACS captured a larger diversity of cell types (Extended Data Fig. 4).

Any individual single-cell sequencing experiment offers only a partial view of cell-type diversity within an organism and gene expression

*A list of authors and their affiliations is available online.

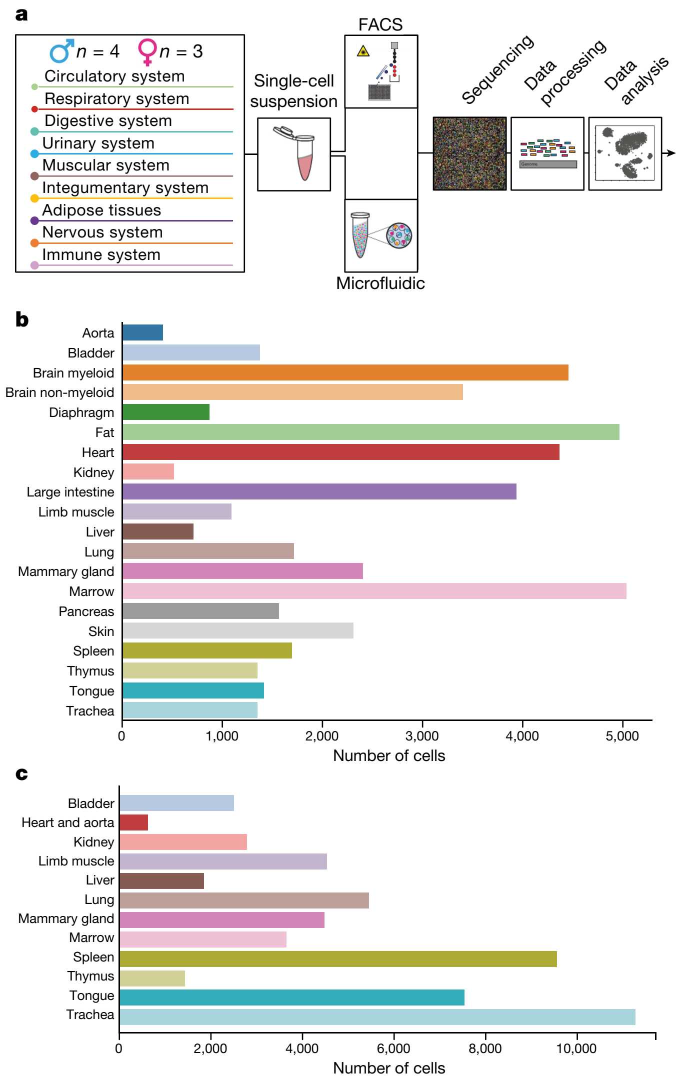


Fig. 1 | Overview of *Tabula Muris*. a, 20 organs from four male and three female mice were analysed. After dissociation, cells were sorted by FACS and, for some organs, captured in microfluidic oil droplets. Cells were lysed, transcriptomes amplified and sequenced, reads mapped, and data analysed. **b**, Bar plot showing the number of sequenced cells prepared by FACS from each organ ($n = 20$ organ types). **c**, Bar plot showing the number of sequenced cells prepared by microfluidic droplets from each organ ($n = 12$ organ types).

within each cell type. We illustrate the expected variability between methods and experiments by comparing our two measurement approaches to a third method, microwell-seq[20]. One notable feature is the variability in the number of genes detected per cell between organs and methods. For example, the median number of genes detected in the bladder is around 4,900 (FACS), 2,900 (microfluidic droplet) and 900 (microwell-seq), whereas in the kidney it is around 1,400 (FACS), 1,900 (microfluidic droplet) and 500 (microwell-seq). In the bladder, liver, lung, mammary gland, trachea, tongue and spleen, nearly twice as many genes are detected per cell with the FACS method compared to the microfluidic-droplet method, whereas the heart and marrow show comparable numbers between the two methods (Extended Data Fig. 5a). This difference is probably not due to sequencing depth, as both FACS and microfluidic-droplet libraries are nearly saturated (Extended Data Fig. 5b). In these comparisons, a gene is considered detected if a single read maps to it, as that is the only value at which reads and UMIs can be treated equally. We also found that the number of detected genes decreases similarly across organs as the read or UMI threshold for a detectable gene is increased (Extended Data Fig. 6).

Next, we investigated whether the three methods agree on the genes defining each cell cluster (Methods). As expected, the FACS and microfluidic-droplet methods show the closest agreement, probably because they used the same biological samples. However, there are several dozen to several hundred genes common to all methods that define each cluster (Extended Data Fig. 7, Supplementary Table 2). This suggests that combining independent datasets can lead to more robust characterizations of gene expression.

Spleen and kidney are two organs for which FACS was performed without marker-based sorting, which enables us to compare the number and relative abundance of different cell types between methods. For those cell types that are captured by both methods, the proportion of each cell type is equivalent (Pearson correlation coefficient: spleen, 0.99; kidney, 0.99). Nonetheless, the microfluidic-droplet method identified cell types that were missed by the FACS method in both organs, for example kidney mesangial cells, and splenic dendritic and natural killer cells. This is partially explained by cellular abundance and sampling depth (12,333 microfluidic-droplet cells compared with 2,216 FACS cells, Supplementary Table 1), and possibly from cell capture and lysis biases between methods.

As the FACS method captures fewer cells but detects more molecules per cell than the microfluidic-droplet method, we asked whether the two methods agree in their 'bulk' gene-expression profiles for the 33 shared cell populations (Methods). Such gene-expression profiles largely correlate (Pearson correlation coefficient: 0.74–0.90), which suggests that although biases between methods exist, both accurately recapitulate average cell-type gene-expression profiles.

Global clustering across organs

To detect relationships between cells from different organs, we visualized all FACS cells with *t*-SNE and grouped them with unbiased, graph-based clustering (Fig. 2, Extended Data Fig. 8). As expected, cells from different organs often mixed, with 25 of 54 clusters containing (at least five) cells from distinct organs (Fig. 3). For example, clusters 3 and 48 each contain endothelial cells from five or more organs, and clusters 1 and 24 contain mesenchymal and stromal cells from four or more organs. Cluster 2 contains B cells from fat, limb muscle, lung, spleen, marrow and liver, but also cells annotated as leukocytes and lymphocytes from the thymus, heart and limb muscle. This suggests that the effect of cell type on measured gene expression is stronger than the effect of batch or dissociation protocol.

Cluster co-membership alone, however, is insufficient to conclude that two cell populations from different organs represent the same or similar cell types; at any given resolution, unbiased clustering that groups related cells may also group unrelated cells[21]. Therefore, to determine which clusters are composed of related or unrelated cell types, we computed a heterogeneity score for each cluster (Methods), and found low scores for the biologically sensible clusters discussed above (Extended Data Fig. 9). By contrast, the astrocytes and epithelial cells in cluster 53 are as different from one another as two random cells.

In addition to these heterogeneous groups, the clustering reveals small populations of potentially mislabelled cells inside homogenous populations. For example, ten thymus cells in cluster 3 (composed of 2,379 cells) are annotated as 'leukocytes', but they express *Pecam1*, which is an endothelial marker. This is a predictable artefact of the annotation scheme: because entire clusters, rather than individual cells, were annotated in each organ, a sufficiently rare cell type that was algorithmically grouped with a more populous cell type will be misannotated. This seems to occur only for populations smaller than about 30 cells, which comprise less than 4% of the overall dataset, and represents the lower limit of sensitivity in the current release of data interpretation.

The fact that most cells of similar cell types cluster together across organs and biological replicates shows that batch effects are not the main source of variance in the dataset. Our findings also show that manual annotation of cell types is consistent with unbiased transcriptomic clustering for sufficiently large populations. We expect that further development of multi-scale comparison algorithms will facilitate

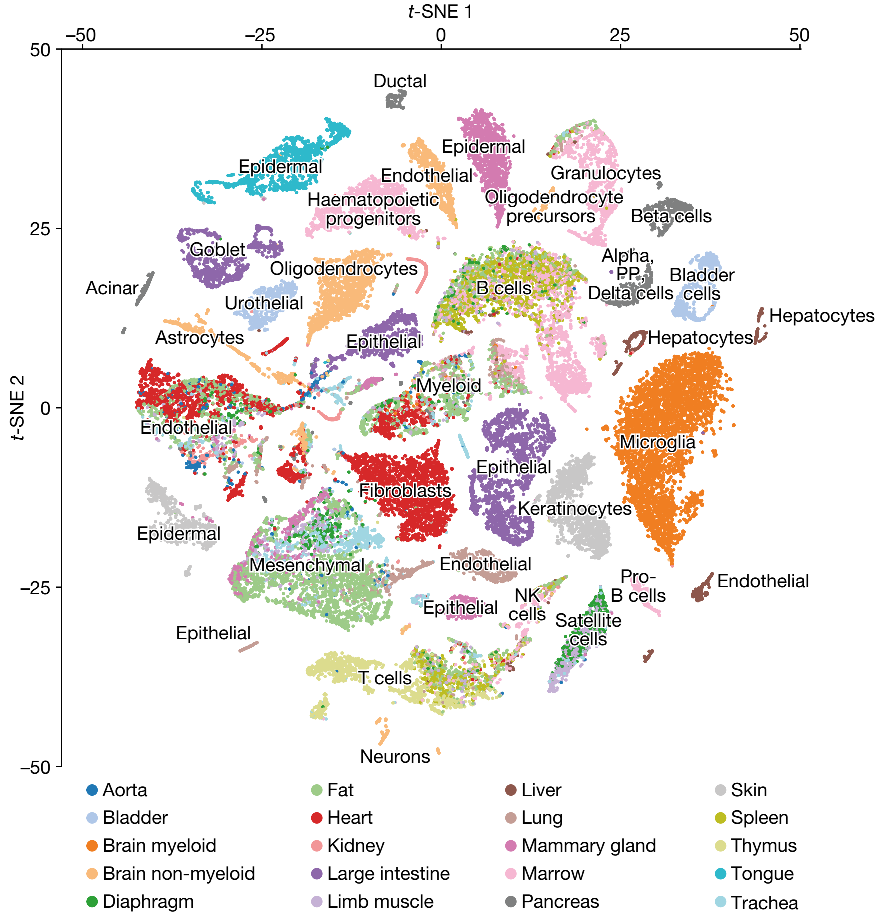


Fig. 2 | *t*-SNE visualization of all FACS cells. *t*-SNE plot of all cells collected by FACS, coloured by organ, overlaid with the predominant cell type composing each cluster; $n = 44{,}949$ individual cells.

the discovery of both universal and organ-specific gene modules within these shared cell types.

To demonstrate an example of investigating common cell types across organs, we collectively analysed all FACS cells annotated as T cells, which revealed five clusters (Fig. 4). Cluster 0 comprises thymic cells undergoing VDJ recombination characterized by the expression of *Rag1*, *Rag2* and *Dntt*, and includes uncommitted double-positive T cells (*Cd4*$^+$ and *Cd8a*$^+$). Cluster 4 contains predominantly proliferating thymic T cells, which may represent pre-T cells expanding after VDJ recombination. Clusters 1–3 contain mostly single-positive T cells (*Cd4*$^+$ or *Cd8a*$^+$). Cluster 3 contains *Cd5*^{hi} thymic T cells that are possibly undergoing positive selection, whereas Cluster 2 contains mostly non-thymic T cells expressing the high-affinity IL2 receptor (encoded by the genes *Il2ra* and *Il2rb*), which suggests that they are activated. Notably, they also express MHC class II genes (*H2-Aa* and *H2-Ab1*). Although this is known in human T cells, MHC class II was previously thought to be restricted to professional antigen-presenting cells in mice[22]. Finally, Cluster 1 also represents mature T cells, but primarily splenic.

Global transcription factor analysis

One major goal of defining cell identities is to understand the underlying regulatory networks. We investigated how transcription factors contribute to cell-type identity by clustering averaged gene-expression profiles for each cell type using only the 1,016 transcription factors expressed in our dataset (Fig. 5a). The resulting dendrogram closely resembles the dendrogram produced using all expressed genes, indicating that transcription factors can be used to reconstruct known cell-ontology relationships between bulk populations (entanglement = 0.11; Extended Data Fig. 10a). By contrast, when we repeated the analysis using cell-surface markers, RNA splicing factors, or the two groups combined (equivalent to a random set of genes), the entanglement was 0.22, 0.25 and 0.34, respectively, which suggests that none of these molecular classes define cell type to the extent that transcription factors do.

We then analysed organ-specific transcription factors by performing correlation analysis on shared cell types between organs[23] (epithelial cells, endothelial cells, B cells and T cells; Fig. 5b–e, Extended Data Fig. 10b–i). To understand which transcription factors were most informative for specifying cell types, we performed variable selection using random forest models (Methods) and determined that 136 transcription factors are needed to simultaneously define all cell types across all organs (Fig. 5f, Supplementary Table 3). We then determined the transcription factor sets that distinguish each individual cell type from all other cells. These sets vary substantially in size (from 2 to 813 transcription factors) and are not necessarily unique to each cell type (Fig. 5g–i, Supplementary Table 4).

A possible application for such transcription factor networks is the design of reprogramming protocols. Indeed, the transcription factors used in published methods are found in the cell-type-specific transcription factors sets we discovered (Supplementary Table 5). For some cell types, such as hepatocytes, satellite cells and oligodendrocytes, those reprogramming factors are the top variables segregating cell types (Fig. 5g–i). In fact, for nearly all reprogramming protocols the transcription factors used also specified the targeted cell type in our data (Supplementary Table 5), which suggests that our data can inform novel reprogramming schemes.

Discussion

A key challenge for single-cell studies is to understand transcriptomic changes caused by dissociation. A previous study showed that quiescent limb-muscle satellite cells activate upon dissociation and consequently

Fig. 3 | Comparison of cell-type determination. Comparison of cell-type determination as performed by unbiased whole-transcriptome comparison versus manual annotation of clusters by organ-specific experts. The x axis represents clusters from Fig. 2, while the y axis represents manual expert annotation of clusters derived from individual organs analysed independently (Extended Data Fig. 1). The unbiased method discovers relationships between similar cell types found in different organs; in particular T cells from different organs are grouped into a single cluster, B cells from different organs into a different single cluster, and endothelial cells from different organs into a single cluster (regions outlined in blue boxes).

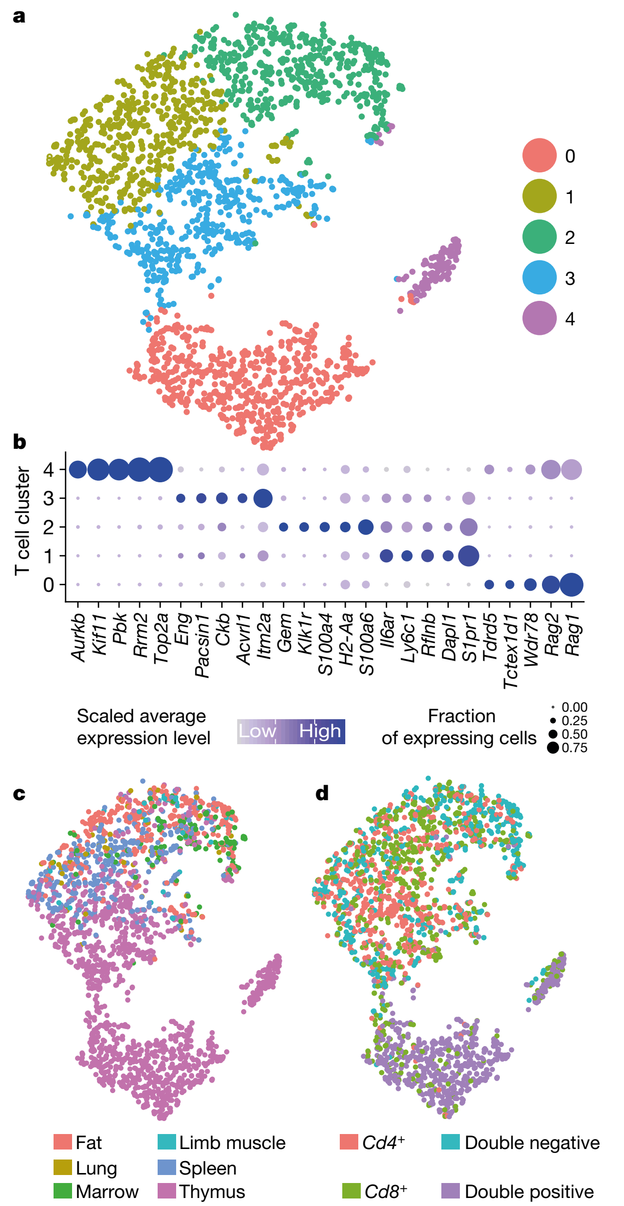


Fig. 4 | Analysis of all sorted T cells. a, *t*-SNE plot of all T cells coloured by cluster membership, highlighting the five identified clusters; $n = 2,847$ individual cells. **b**, Dot plot showing level of expression (colour scale) and number of expressing cells (point diameter) within each cluster of T cells. *Rflnb* is also known as *Fam101b*. **c**, *t*-SNE plot of all T cells coloured by organ of origin (fat, lung, marrow, limb muscle, spleen or thymus); $n = 2,847$ individual cells. **d**, *t*-SNE plot of all T cells coloured by classification of T cells to four categories based on expression of *Cd4* and *Cd8* (*Cd4*$^+$, *Cd8*$^+$, *Cd4*$^+$*Cd8*$^+$, *Cd4*$^-$*Cd8*$^-$); $n = 2,847$ individual cells.

express immediate early genes and other dissociation-related markers[24]. We clearly observed these markers in several organs including limb muscle (Extended Data Fig. 11), but many showed little evidence of cellular activation. Therefore, the dissociation-related satellite-cell markers are not universal, and organs probably display unique dissociation-related expression profiles. Importantly, the presence of such changes in gene expression does not prevent the identification of cell type or the comparison of cell types across organs.

Another challenge for single-cell studies is experimental design amid the choice of several technologies. Droplet-based technologies offer certain advantages in the discovery of rare cell types or states, for example when many cells (tens of thousands) are required to reconstruct whole-organism architecture and developmental lineages[25,26]. FACS-based methods generate high coverage over small cellular populations (tens to thousands), and are beneficial for enriching specific or rare cell types, and for studying subtle heterogeneity involving lowly expressed genes[27], alternative splicing[15] and sequence variation analysis[28]. There are opportunities to combine the two methods, such as by running sorted cells on a microfluidic-droplet platform, which could potentially accommodate both cell-type enrichment and cost factors.

Recently, a complementary scRNA-seq study across mouse organs was published[20]. Those data contained four times as many cells and included several sample types not present in our data, such as neonatal and fetal organs, cell lines, and young adult ovary, peripheral blood,

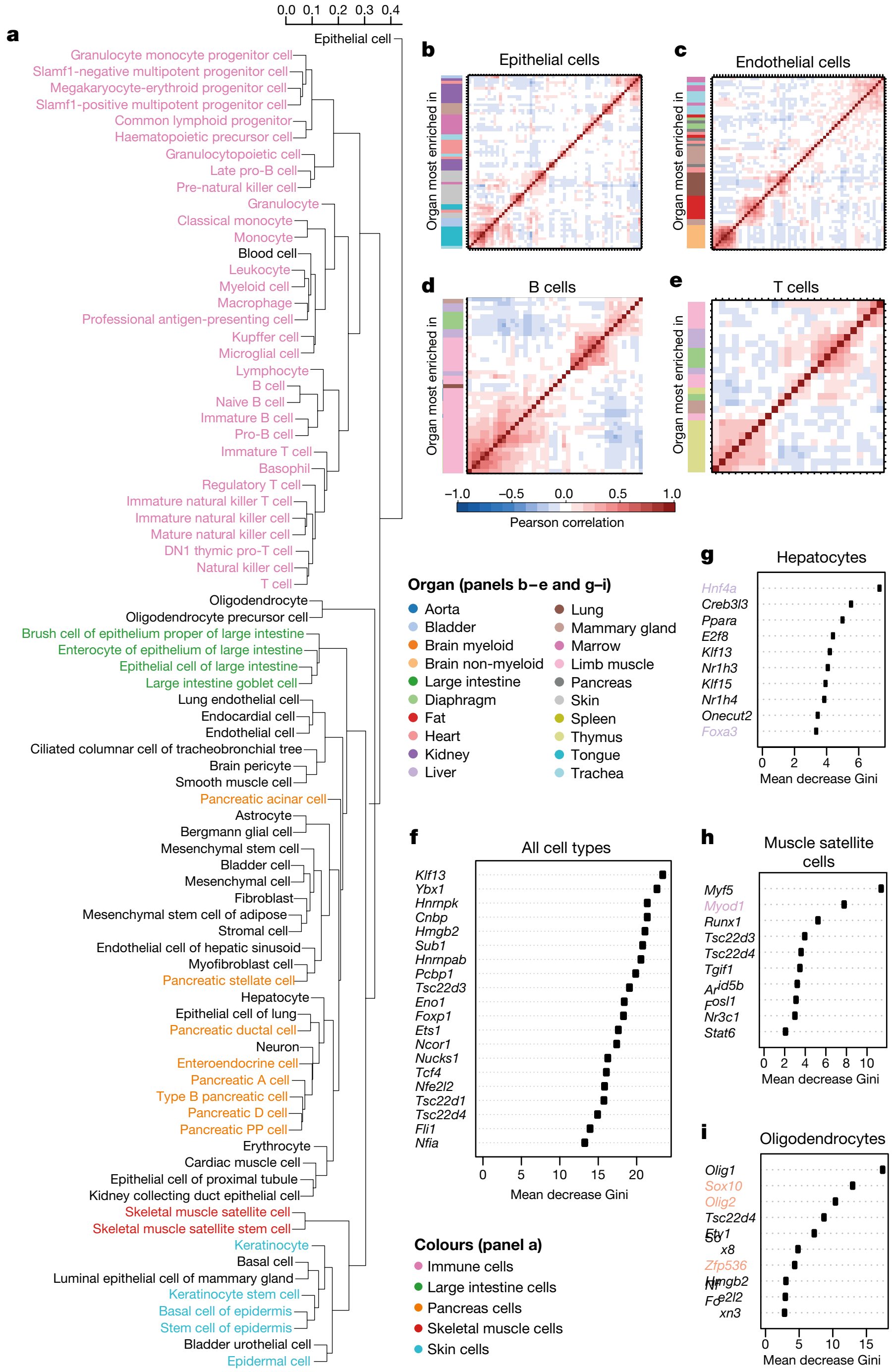


Fig. 5 | Transcription factor analysis. a, Dendrogram of cell types constructed with only transcription factors. **b**–**e**, Correlograms of top organ-specific transcription factors for epithelial cells (**b**), endothelial cells (**c**), B cells (**d**) and T cells (**e**). Row colours correspond to the organ of the most-enriched cell type; $n = 60$ randomly selected cells for each cell type. **f**, Top 20 transcription factors (mean Gini importance) of the random-forest model when classifying all cell types. **g**–**i**, Top 10 transcription factors (mean Gini importance) of the random-forest model when classifying each cell type individually. The coloured genes correspond to transcription factors used in successful reprogramming protocols.

placenta, prostate, small intestine, stomach, testis and uterus. However, our FACS data contain four times as many genes per cell, and we analysed several organs not present in the other dataset[20], such as aorta, four brain regions, diaphragm, four fat types, four adult heart chambers, adult telogen and anagen skin, tongue and trachea. Additionally, several features of our study facilitate replication and cross-experiment analysis: all data, analysis and code are freely available; our web portal enables one to query gene expression in all organs simultaneously; we annotated cell types using standard cell ontology terms, thereby enabling cross-organ and cross-experiment analyses; age and sex are controlled in our data by collecting all organs from the same mice; both sexes are represented for all organs in our data; organs were perfused, enabling the analysis of tissue-resident immune cells; and full-length transcript data make possible transcription factor, splice variant, and sequence variant analyses.

In conclusion, we have created a compendium of single-cell transcriptional measurements across 20 mouse organs. This *Tabula Muris*, or 'Mouse Atlas', has many uses, including the discovery of new putative cell types, the discovery of novel gene expression in known cell types, and the ability to compare cell types across organs. It will also serve as a reference of healthy young adult organs, which can be used as a baseline for current and future mouse models of disease. Although

it is not an exhaustive characterization of all mouse organs, it does provide a rich dataset of the most highly studied organs in biology. The *Tabula Muris* provides a framework and description of many of the most populous and important cell populations within the mouse, and represents a foundation for future studies across a multitude of diverse physiological disciplines.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0590-4

Received: 20 December 2017; Accepted: 20 August 2018;
Published online: 03 October 2018

1. Alberts, B. et al. *Essential Cell Biology* (W.W. Norton & Company, New York, 2016).
2. Guo, G. et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* **18**, 675–685 (2010).
3. Dalerba, P. et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol*. **29**, 1120–1127 (2011).
4. Thorsen, T., Roberts, R. W., Arnold, F. H. & Quake, S. R. Dynamic pattern formation in a vesicle-generating microfluidic device. *Phys. Rev. Lett.* **86**, 4163–4166 (2001).
5. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
6. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
7. Ramsköld, D. et al. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol*. **30**, 777–782 (2012).
8. Wu, A. R. et al. Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
9. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
10. Enge, M. et al. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* **171**, 321–330.e14 (2017).
11. Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).
12. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
13. Villani, A.-C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
14. Darmanis, S. et al. A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl Acad. Sci. USA* **112**, 7285–7290 (2015).
15. Gokce, O. et al. Cellular taxonomy of the mouse striatum as revealed by single-cell RNA-seq. *Cell Rep.* **16**, 1126–1137 (2016).
16. Usoskin, D. et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci*. **18**, 145–153 (2015).
17. Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
18. Li, H. et al. Classifying *Drosophila* olfactory projection neuron subtypes by single-cell RNA sequencing. *Cell* **171**, 1206–1220.e22 (2017).
19. Bakken, T. et al. Cell type discovery and representation in the era of high-content single cell phenotyping. *BMC Bioinformatics* **18** (Suppl 17), 559 (2017).
20. Han, X. et al. Mapping the mouse cell atlas by microwell-seq. *Cell* **172**, 1091–1107.e17 (2018).
21. Freytag, S., Tian, L., Lonnstedt, I., Ng, M. & Bahlo, M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Res* **7**, 1297 (2018).
22. Holling, T. M., Schooten, E. & van Den Elsen, P. J. Function and regulation of MHC class II molecules in T-lymphocytes: of mice and men. *Hum. Immunol*. **65**, 282–290 (2004).
23. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E* **74**, 016110 (2006).
24. van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
25. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
26. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
27. Liu, Z. et al. Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. *Nature* **551**, 100–104 (2017).
28. Darmanis, S. et al. Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep*. **21**, 1399–1410 (2017).

Acknowledgements We thank Sony Biotechnology for making an SH800S instrument available for this project. Some of the cell sorting/flow cytometry analysis for this project was performed using a Sony SH800S instrument in the Stanford Shared FACS Facility. Some FACS experiments used instruments in the VA Flow Cytometry Core, which is supported by the US Department of Veterans Affairs, Palo Alto Veterans Institute for Research and the National Institutes of Health. This work was supported by the Chan Zuckerberg Biohub, NIH Grant DP1 AG053015 and the NOMIS Foundation (T.W.-C.) as well as partly by the Stanford Islet Research Core in the Stanford Diabetes Research Center (P30 DK116074). We thank A. McGeever for contributions to the design of the *Tabula Muris* web portal.

Author contributions See author list for full contributions.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0590-4.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0590-4.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.R.Q., T.W.-C. or S.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The Tabula Muris Consortium

Overall coordination: Nicholas Schaum[1], Jim Karkanias[2], Norma F. Neff[2], Andrew P. May[2], Stephen R. Quake[2,3*], Tony Wyss-Coray[4,5,6*] & Spyros Darmanis[2*]

Logistical coordination: Joshua Batson[2], Olga Botvinnik[2], Michelle B. Chen[3], Steven Chen[2], Foad Green[2], Robert C. Jones[3], Ashley Maynard[2], Lolita Penland[2], Angela Oliveira Pisco[2], Rene V. Sit[2], Geoffrey M. Stanley[3], James T. Webber[2] & Fabio Zanini[3]

Organ collection and processing: Ankit S. Baghel[1], Isaac Bakerman[1,7,8], Ishita Bansal[2], Daniela Berdnik[4], Biter Bilen[4], Douglas Brownfield[9], Corey Cain[10], Michelle B. Chen[3], Steven Chen[2], Min Cho[2], Giana Cirolia[2], Stephanie D. Conley[1], Spyros Darmanis[2], Aaron Demers[2], Kubilay Demir[1,11], Antoine de Morree[4], Tessa Divita[2], Haley du Bois[4], Laughing Bear Torrez Dulgeroff[1], Hamid Ebadi[2], F. Hernán Espinoza[9], Matt Fish[1,11,12], Qiang Gan[4], Benson M. George[1], Astrid Gillich[9], Foad Green[2], Geraldine Genetiano[2], Xueying Gu[12], Gunsagar S. Gulati[1], Yan Hang[12], Shayan Hosseinzadeh[2], Albin Huang[4], Tal Iram[4], Taichi Isobe[1], Feather Ives[2], Robert C. Jones[3], Kevin S. Kao[1], Guruswamy Karnam[13], Aaron M. Kershner[1], Bernhard M. Kiss[1,14], William Kong[1], Maya E. Kumar[15,16], Jonathan Y. Lam[12], Davis P. Lee[6], Song E. Lee[4], Guang Li[17], Qingyun Li[18], Ling Liu[4], Annie Lo[2], Wan-Jin Lu[1,9], Anoop Manjunath[1], Andrew P. May[2], Kaia L. May[2], Oliver L. May[2], Ashley Maynard[2], Marina McKay[2], Ross J. Metzger[19,20], Marco Mignardi[3], Dullei Min[21], Ahmad N. Nabhan[9], Norma F. Neff[2], Katharine M. Ng[3], Joseph Noh[1], Rasika Patkar[13], Weng Chuan Peng[12], Lolita Penland[2], Robert Puccinelli[2], Eric J. Rulifson[12], Nicholas Schaum[1], Shaheen S. Sikandar[1], Rahul Sinha[1,22,23,24], Rene V. Sit[2], Krzysztof Szade[1,25], Weilun Tan[2], Cristina Tato[2], Krissie Tellez[12], Kyle J. Travaglini[9], Carolina Tropini[26], Lucas Waldburger[2], Linda J. van Weele[1], Michael N. Wosczyna[4], Jinyi Xiang[1], Soso Xue[3], Justin Youngyunpipatkul[2], Fabio Zanini[3], Macy E. Zardeneta[6], Fan Zhang[19,20] & Lu Zhou[18]

Library preparation and sequencing: Ishita Bansal[2], Steven Chen[2], Min Cho[2], Giana Cirolia[2], Spyros Darmanis[2], Aaron Demers[2], Tessa Divita[2], Hamid Ebadi[2], Geraldine Genetiano[2], Foad Green[2], Shayan Hosseinzadeh[2], Feather Ives[2], Annie Lo[2], Andrew P. May[2], Ashley Maynard[2], Marina McKay[2], Norma F. Neff[2], Lolita Penland[2], Rene V. Sit[2], Weilun Tan[2], Lucas Waldburger[2] & Justin Youngyunpipatkul[2]

Computational data analysis: Joshua Batson[2], Olga Botvinnik[2], Paola Castro[2], Derek Croote[3], Spyros Darmanis[2], Joseph L. DeRisi[2,27], Jim Karkanias[2], Angela Oliveira Pisco[2], Geoffrey M. Stanley[3], James T. Webber[2] & Fabio Zanini[3]

Cell type annotation: Ankit S. Baghel[1], Isaac Bakerman[1,7,8], Joshua Batson[2], Biter Bilen[4], Olga Botvinnik[2], Douglas Brownfield[9], Michelle B. Chen[3], Spyros Darmanis[2], Kubilay Demir[1,11], Antoine de Morree[4], Hamid Ebadi[2], F. Hernán Espinoza[9], Matt Fish[1,11,12], Qiang Gan[4], Benson M. George[1], Astrid Gillich[9], Xueying Gu[12], Gunsagar S. Gulati[1], Yan Hang[12], Albin Huang[4], Tal Iram[4], Taichi Isobe[1], Guruswamy Karnam[13], Aaron M. Kershner[1], Bernhard M. Kiss[1,14], William Kong[1], Christin S. Kuo[9,11,21], Jonathan Y. Lam[12], Benoit Lehallier[4], Guang Li[17], Qingyun Li[18], Ling Liu[4], Wan-Jin Lu[1,9], Dullei Min[21], Ahmad N. Nabhan[9], Katharine M. Ng[3], Patricia K. Nguyen[1,7,8,17], Rasika Patkar[13], Weng Chuan Peng[12], Lolita Penland[2], Eric J. Rulifson[12], Nicholas Schaum[1], Shaheen S. Sikandar[1], Rahul Sinha[1,22,23,24], Krzysztof Szade[1,25], Serena Y. Tan[22], Krissie Tellez[12], Kyle J. Travaglini[9], Carolina Tropini[26], Linda J. van Weele[1], Bruce M. Wang[13], Michael N. Wosczyna[4], Jinyi Xiang[1], Hanadie Yousef[4] & Lu Zhou[18]

Writing group: Joshua Batson[2], Olga Botvinnik[2], Steven Chen[2], Spyros Darmanis[2], Foad Green[2], Andrew P. May[2], Ashley Maynard[2], Angela Oliveira Pisco[2], Stephen R. Quake[2,3], Nicholas Schaum[1], Geoffrey M. Stanley[3], James T. Webber[2], Tony Wyss-Coray[4,5,6] & Fabio Zanini[3]

Supplemental text writing group: Philip A. Beachy[1,9,11,12], Charles K. F. Chan[28], Antoine de Morree[4], Benson M. George[1], Gunsagar S. Gulati[1], Yan Hang[12], Kerwyn Casey Huang[2,3,26], Tal Iram[4], Taichi Isobe[1], Aaron M. Kershner[1], Bernhard M. Kiss[1,14], William Kong[1], Guang Li[17], Qingyun Li[18], Ling Liu[4], Wan-Jin Lu[1,9], Ahmad N. Nabhan[9], Katharine M. Ng[3], Patricia K. Nguyen[1,7,8,17], Weng Chuan Peng[12], Eric J. Rulifson[12], Nicholas Schaum[1], Shaheen S. Sikandar[1], Rahul Sinha[1,22,23,24], Krzysztof Szade[1,25], Kyle J. Travaglini[9], Carolina Tropini[26], Bruce M. Wang[13], Kenneth Weinberg[21], Michael N. Wosczyna[4], Sean M. Wu[17] & Hanadie Yousef[4]

Principal investigators: Ben A. Barres[18], Philip A. Beachy[1,9,11,12], Charles K. F. Chan[28], Michael F. Clarke[1], Spyros Darmanis[2], Kerwyn Casey Huang[2,3,26], Jim Karkanias[2], Seung K. Kim[12,29], Mark A. Krasnow[9,11], Maya E. Kumar[15,16], Christin S. Kuo[9,11,21], Andrew P. May[2], Ross J. Metzger[19,20], Norma F. Neff[2], Roel Nusse[9,11,12], Patricia K. Nguyen[1,7,8,17], Thomas A. Rando[4,5,6], Justin Sonnenburg[2,26], Bruce M. Wang[13], Kenneth Weinberg[21], Irving L. Weissman[1,22,23,24], Sean M. Wu[1,7,17], Stephen R. Quake[2,3] & Tony Wyss-Coray[4,5,6]

[1]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA. [2]Chan Zuckerberg Biohub, San Francisco, CA, USA. [3]Department of Bioengineering, Stanford University, Stanford, CA, USA. [4]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA, USA. [5]Paul F. Glenn Center for the Biology of Aging, Stanford University School of Medicine, Stanford, CA, USA. [6]Center for Tissue Regeneration, Repair, and Restoration, VA Palo Alto Healthcare System, Palo Alto, CA, USA. [7]Stanford Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA, USA. [8]Department of Medicine, Division of Cardiology, Stanford University School of Medicine, Stanford, CA, USA. [9]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, USA. [10]Flow Cytometry Core, VA Palo Alto Healthcare System, Palo Alto, CA, USA. [11]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. [12]Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA, USA. [13]Department of Medicine and Liver Center, University of California San Francisco, San Francisco, CA, USA. [14]Department of Urology, Stanford University School of Medicine, Stanford, CA, USA. [15]Sean N. Parker Center for Asthma and Allergy Research, Stanford University School of Medicine, Stanford, CA, USA. [16]Department of Medicine, Division of Pulmonary and Critical Care, Stanford University School of Medicine, Stanford, CA, USA. [17]Department of Medicine, Division of Cardiovascular Medicine, Stanford University, Stanford, CA, USA. [18]Department of Neurobiology, Stanford University School of Medicine, Stanford, CA, USA. [19]Vera Moulton Wall Center for Pulmonary and Vascular Disease, Stanford University School of Medicine, Stanford, CA, USA. [20]Department of Pediatrics, Division of Cardiology, Stanford University School of Medicine, Stanford, CA, USA. [21]Department of Pediatrics, Pulmonary Medicine, Stanford University School of Medicine, Stanford, CA, USA. [22]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [23]Ludwig Center for Cancer Stem Cell Research and Medicine, Stanford University School of Medicine, Stanford, CA, USA. [24]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. [25]Department of Medical Biotechnology, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Kraków, Poland. [26]Department of Microbiology & Immunology, Stanford University School of Medicine, Stanford, CA, USA. [27]Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA. [28]Department of Surgery, Division of Plastic and Reconstructive Surgery, Stanford University, Stanford, CA, USA. [29]Department of Medicine and Stanford Diabetes Research Center, Stanford University, Stanford, CA, USA. *e-mail: quake@stanford.edu; twc@stanford.edu; spyros.darmanis@czbiohub.org

METHODS

Mice and organ collection. Four 10–15 week old male and four virgin female C57BL/6JN mice were shipped from the National Institute on Aging colony at Charles River (housed at 67–73 °F) to the Veterinary Medical Unit (VMU; housed at 68–76 °F)) at the VA Palo Alto (VA). At both locations, mice were housed on a 12-h light/dark cycle, and provided food and water ad libitum. The diet at Charles River was NIH-31, and Teklad 2918 at the VA VMU. Littermates were not recorded or tracked, and mice were housed at the VA VMU for no longer than 2 weeks before euthanasia. Before tissue collection, mice were placed in sterile collection chambers at 8 am for 15 min to collect fresh fecal pellets. After anaesthetization with 2.5% v/v Avertin, mice were weighed, shaved, and blood was drawn via cardiac puncture before transcardial perfusion with 20 ml PBS. Mesenteric adipose tissue was then immediately collected to avoid exposure to the liver and pancreas perfusate, which negatively affects cell sorting. Isolating viable single cells from both the pancreas and the liver of the same mouse was not possible; therefore, two males and two females were used for each. Whole organs were then dissected in the following order: large intestine, spleen, thymus, trachea, tongue, brain, heart, lung, kidney, gonadal adipose tissue, bladder, diaphragm, limb muscle (tibialis anterior), skin (dorsal), subcutaneous adipose tissue (inguinal pad), mammary glands (fat pads 2, 3 and 4), brown adipose tissue (interscapular pad), aorta and bone marrow (spine and limb bones). Organ collection concluded by 10 am. After single-cell dissociation as described below, cell suspensions were either used for FACS of individual cells into 384-well plates, or for preparation of the microfluidic droplet library. All animal care and procedures were carried out in accordance with institutional guidelines approved by the VA Palo Alto Committee on Animal Research.

Tissue dissociation and sample preparation. Specific protocols for each tissue are described in the Supplementary Information.

Sample size, randomization and blinding. No sample size choice was performed before the study. Randomization and blinding were not performed: the authors were aware of all data and metadata-related variables during the entire course of the study.

Single-cell methods. *Lysis plate preparation.* Lysis plates were created by dispensing 0.4 μl lysis buffer (0.5 U Recombinant RNase Inhibitor (Takara Bio, 2313B), 0.0625% Triton™ X-100 (Sigma, 93443-100ML), 3.125 mM dNTP mix (Thermo Fisher, R0193), 3.125 μM Oligo-dT$_{30}$VN (Integrated DNA Technologies, 5′AAGCAGTGGTATCAACGCAGAGTACT$_{30}$VN-3′) and 1:600,000 ERCC RNA spike-in mix (Thermo Fisher, 4456740)) into 384-well hard-shell PCR plates (Bio-Rad HSP3901) using a Tempest liquid handler (Formulatrix). 96-well lysis plates were also prepared with 4 μl lysis buffer. All plates were sealed with AlumaSeal CS Films (Sigma-Aldrich Z722634) and spun down (3,220*g*, 1 min) and snap-frozen on dry ice. Plates were stored at −80 °C until sorting.

FACS. After dissociation, single cells from each organ and tissue were isolated into 384- or 96-well plates via FACS. Most organs were sorted into 384-well plates using SH800S (Sony) sorters. Heart and liver were sorted into 96-well plates and cardiomyocytes were hand-picked into 96-well plates. Limb muscle and diaphragm were sorted into 384-well plates on an Aria III (Becton Dickinson) sorter. The last two columns of each 384 well plate were intentionally left as blanks. For most organs, single cells were selected with forward scatter, and dead cells and common cell types were excluded with a single colour channel. Combinations of fluorescent antibodies were used for most organs to enrich for rare cell populations (see Supplementary Information), but some were stained only for viable cells. Colour compensation was used whenever necessary. On the SH800, the highest purity setting (‘Single cell’) was used for all but the rarest cell types, for which the ‘Ultrapure’ setting was used. Sorters were calibrated using FACS buffer every day before collecting any cells, and also after every eight sorted plates. For a typical sort, 1–3 ml of pre-stained cell suspension was filtered, vortexed gently, and loaded onto the FACS machine. A small number of cells were flowed at low pressure to check cell and debris concentrations. The pressure was then adjusted, flow paused, the first destination plate unsealed and loaded, and sorting started. If a cell suspension was too concentrated, it was diluted using FACS buffer or 1X PBS. For some cell types, such as hepatocytes, 96-well plates were used because it was not possible to sort individual cells accurately into 384-well plates. Immediately after sorting, plates were sealed with a pre-labelled aluminium seal, centrifuged, and flash frozen on dry ice. On average, each 384-well plate took 8 min to sort.

cDNA synthesis and library preparation. cDNA synthesis was performed using the Smart-seq2 protocol[7,8]. In brief, 384-well plates containing single-cell lysates were thawed on ice followed by first-strand synthesis. 0.6 μl of reaction mix (16.7 U μl^{-1} SMARTScribe Reverse Transcriptase (Takara Bio, 639538), 1.67 U μl^{-1} Recombinant RNase Inhibitor (Takara Bio, 2313B), 1.67X First-Strand Buffer (Takara Bio, 639538), 1.67 μM TSO (Exiqon, 5′-AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG-3′), 8.33 mM dithiothreitol (Bioworld, 40420001-1), 1.67 M Betaine (Sigma, B0300-5VL) and 10 mM $MgCl_2$ (Sigma, M1028-10X1ML)) was added to each well using a Tempest liquid handler. Reverse transcription was carried out by incubating wells on a ProFlex 2 × 384 thermal-cycler (Thermo Fisher) at 42 °C for 90 min, and stopped by heating at 70 °C for 5 min.

Subsequently, 1.5 μl of PCR mix (1.67X KAPA HiFi HotStart ReadyMix (Kapa Biosystems, KK2602), 0.17 μM IS PCR primer (IDT, 5′-AAGCAGTGGTAT CAACGCAGAGT-3′), and 0.038 U μl^{-1} Lambda Exonuclease (NEB, M0262L)) was added to each well with a Mantis liquid handler (Formulatrix), and second-strand synthesis was performed on a ProFlex 2x384 thermal-cycler by using the following program: 1) 37 °C for 30 min, 2) 95 °C for 3 min, 3) 23 cycles of 98 °C for 20 s, 67 °C for 15 s and 72 °C for 4 min, and 4) 72 °C for 5 min.

The amplified product was diluted with a ratio of 1 part cDNA to 10 parts 10 mM Tris-HCl (Thermo Fisher, 15568025), and concentrations were measured with a dye-fluorescence assay (Quant-iT dsDNA High Sensitivity kit; Thermo Fisher, Q33120) on a SpectraMax i3x microplate reader (Molecular Devices). Sample plates were selected for downstream processing if the mean concentration of blanks (ERCC-containing, non-cell wells) was greater than 0 ng μl^{-1}, and, after linear regression of the values obtained from the Quant-iT dsDNA standard curve, the R^2 value was greater than 0.98. Sample wells were then selected if their cDNA concentrations were at least one standard deviation greater than the mean concentration of the blanks. These wells were reformatted to a new 384-well plate at a concentration of 0.3 ng μl^{-1} and a final volume of 0.4 μl using an Echo 550 acoustic liquid dispenser (Labcyte).

Illumina sequencing libraries were prepared as described previously[14]. In brief, tagmentation was carried out on double-stranded cDNA using the Nextera XT Library Sample Preparation kit (Illumina, FC-131-1096). Each well was mixed with 0.8 μl Nextera tagmentation DNA buffer (Illumina) and 0.4 μl Tn5 enzyme (Illumina), then incubated at 55 °C for 10 min. The reaction was stopped by adding 0.4 μl Neutralize Tagment Buffer (Illumina) and centrifuging at room temperature at 3,220*g* for 5 min. Indexing PCR reactions were performed by adding 0.4 μl of 5 μM i5 indexing primer, 0.4 μl of 5 μM i7 indexing primer, and 1.2 μl of Nextera NPM mix (Illumina). PCR amplification was carried out on a ProFlex 2x384 thermal cycler using the following program: 1) 72 °C for 3 min, 2) 95 °C for 30 s, 3) 12 cycles of 95 °C for 10 s, 55 °C for 30 s and 72 °C for 1 min, and 4) 72 °C for 5 min.

Library pooling, quality control and sequencing. After library preparation, wells of each library plate were pooled using a Mosquito liquid handler (TTP Labtech). Pooling was followed by two purifications using 0.7x AMPure beads (Fisher, A63881). Library quality was assessed using capillary electrophoresis on a Fragment Analyzer (AATI), and libraries were quantified by qPCR (Kapa Biosystems, KK4923) on a CFX96 Touch Real-Time PCR Detection System (Biorad). Plate pools were normalized to 2 nM and equal volumes from 10 or 20 plates were mixed together to make the sequencing sample pool. A PhiX control library was spiked in at 0.2% before sequencing.

Sequencing libraries from 384-well and 96-well plates. Libraries were sequenced on the NovaSeq 6000 Sequencing System (Illumina) using 2 × 100-bp paired-end reads and 2 × 8-bp or 2 × 12-bp index reads with either a 200- or 300-cycle kit (Illumina, 20012861 or 20012860).

Microfluidic droplet single-cell analysis. Single cells were captured in droplet emulsions using the GemCode Single-Cell Instrument (10x Genomics), and scRNA-seq libraries were constructed as per the 10x Genomics protocol using GemCode Single-Cell 3′ Gel Bead and Library V2 Kit. In brief, single cell suspensions were examined using an inverted microscope, and if sample quality was deemed satisfactory, the sample was diluted in PBS with 2% FBS to a concentration of 1000 cells per μl. If cell suspensions contained cell aggregates or debris, two additional washes in PBS with 2% FBS at 300*g* for 5 min at 4 °C were performed. Cell concentration was measured either with a Moxi GO II (Orflo Technologies) or a haemocytometer. Cells were loaded in each channel with a target output of 5,000 cells per sample. All reactions were performed in the Biorad C1000 Touch Thermal cycler with 96-Deep Well Reaction Module. 12 cycles were used for cDNA amplification and sample index PCR. Amplified cDNA and final libraries were evaluated on a Fragment Analyzer using a High Sensitivity NGS Analysis Kit (Advanced Analytical). The average fragment length of 10x cDNA libraries was quantitated on a Fragment Analyzer (AATI), and by qPCR with the Kapa Library Quantification kit for Illumina. Each library was diluted to 2 nM, and equal volumes of 16 libraries were pooled for each NovaSeq sequencing run. Pools were sequenced with 100 cycle run kits with 26 bases for Read 1, 8 bases for Index 1, and 90 bases for Read 2 (Illumina 20012862). A PhiX control library was spiked in at 0.2 to 1%. Libraries were sequenced on the NovaSeq 6000 Sequencing System (Illumina).

Data processing. Sequences from the NovaSeq were de-multiplexed using bcl-2fastq version 2.19.0.316. Reads were aligned using to the mm10plus genome using STAR version 2.5.2b with parameters TK. Gene counts were produced using HTSEQ version 0.6.1p1 with default parameters, except ‘stranded’ was set to ‘false’, and ‘mode’ was set to ‘intersection-nonempty’.

Sequences from the microfluidic droplet platform were de-multiplexed and aligned using CellRanger version 2.0.1, available from 10x Genomics with default parameters.

Clustering. Standard procedures for filtering, variable gene selection, dimensionality reduction and clustering were performed using the Seurat package version 2.2.1. A detailed worked example, including the mathematical formulae for each operation, is in the Organ Annotation Vignette. The parameters that were tuned on a per-tissue basis (resolution and number of principal components (PCs)) can be viewed in the tissue-specific Rmd files available on GitHub. For each tissue and each sequencing method (FACS and microfluidic droplet), the following steps were performed:

1. Cells were lexicographically sorted by cell ID to ensure reproducibility.
2. Cells with fewer than 500 detected genes were excluded. (A gene counts as detected if it has at least one read mapping to it). Cells with fewer than 50,000 reads (FACS) or 1,000 UMI (microfluidic droplet) were excluded.
3. Counts were log-normalized for each cell using the natural logarithm of 1 + counts per million (for FACS) or 1 + counts per ten thousand (for microfluidic droplet).
4. Variable genes were selected using a threshold (0.5) for the standardized log dispersion, in which the standardization was performed separately according to binned values of log mean expression.
5. The variable genes were projected onto a low-dimensional subspace using principal component analysis. The number of principal components was selected on the basis of inspection of the plot of variance explained.
6. A shared-nearest-neighbours graph was constructed on the basis of the Euclidean distance in the low-dimensional subspace spanned by the top principal components. Cells were clustered using a variant of the Louvain method that includes a resolution parameter in the modularity function[13].
7. Cells were visualized using a 2-dimensional t-distributed Stochastic Neighbour Embedding of the PC-projected data.
8. Cell types were assigned to each cluster using the abundance of known marker genes. Plots showing the expression of the markers for each tissue appear in the Extended Data.
9. When clusters appeared to be mixtures of cell types, they were refined either by increasing the resolution parameter for clustering or subsetting the data and rerunning steps 3–7.

A similar analysis was done globally for all FACS-processed cells and for all microfluidic-droplet-processed cells to produce an unbiased clustering.

Heterogeneity score. Let C be a cluster, decomposed into annotated cell types $C = T_1 \cup \cdots \cup T_k$. For each pair of cell types T_i, T_j, we compute the average distance between their members: $d_{ij} = \frac{1}{|T_i||T_j|}\sum_{x\in T_i, y\in T_j}|x-y|$. The heterogeneity score C is the maximum of those distances over cell types T with at least five cells. For the FACS data, the vector x for a cell is the PC-projection from step 5 above. Extended Data Fig. 9 contains heat maps of the cell-type distance matrix d_{ij} for select clusters and a bar plot of the heterogeneity scores for all clusters containing several cell types.

Differential expression overlap analysis. For FACS and microfluidic droplet data, differential expression analysis for each organ was performed using a Wilcoxon rank-sum test as implemented in the 'FindAllMarkers' function of the Seurat package. Differential expression was performed between cell ontology groups and resulted in a list of differentially expressed genes (ln(FoldChange) >0.25) between each cell ontology group and all other ontology groups of the same organ. For microwell-seq we used the corresponding published lists for each cell type and for every organ. We then assessed the overlap of those lists between the three methods. As the nomenclature is not identical, the analysis was performed between cell types that could be matched with a certain degree of confidence between the three methods (Supplementary Table 2).

Correlating bulk gene expression profiles. For the 33 cell populations shared between FACS and microfluidic droplets, the average gene-expression profile of each population was calculated. The quality of such a bulk gene-expression profile depends on the total number of detected molecules. FACS detects more molecules per cell, but fewer cells. Microfluidic droplets detect fewer molecules per cell, but more cells. To assess the agreement between methods on annotated cell types, Pearson correlation was used on the log expression profiles of each shared cell population. (Only genes present at 1 count per million or greater in at least one of the datasets were considered. A pseudocount of 1 count per million was added before taking logarithms.)

Calculation of dissociation scores. For each organ, principal component analysis was performed on a subset of 140 dissociation-related genes[23]. The first principal component was used as the 'dissociation score' as it corresponds to the variance within these genes.

Defining cell type-enriched transcription factors. Transcription factors were defined as the 1,140 genes annotated by the Gene Ontology term 'DNA binding transcription factor activity', downloaded from the Mouse Genome Informatics database (http://www.informatics.jax.org/mgihome/GO/project.shtml, accessed on 10 November 2017. Cell types were defined as unique combinations of cell ontology and organ annotation (for example, Lung_Endothelial_cell). All analyses were performed on the full dataset, except the correlograms for which the data was subsampled by randomly selecting 60 cells from each cell type. Enriched transcription factors were defined by the Seurat FindMarkers function with the Wilcoxon significance test for the target cell type against the all of the rest of the cell types combined. These were filtered by p_val $< 10^{-3}$, avg_diff > 0.2, pct.1 $-$ pct.2 > 0.1 (per cent detected difference > 0.1), and pct.1 > 0.3 (detected in $>$30% of target cells).

Cell-type comparisons between methods using cell ontology classes. We used the OntologyX R package family version 2.4 (libraries ontologyIndex, ontologyPlot, and ontologySimilarity) to draw the representative cell ontology dendrograms (function onto_plot). To compute the tanglegram (function tanglegram from dendextend R package version 1.8) we used the dendrogram created from all expressed genes as the reference for comparisons to the dendrograms produced using particular gene ontology cellular functions (transcription factors, cell surface markers, RNA splicing factors). The entanglement scores were calculated using the step2side method (function untangle from dendextend R package). Entanglement is a measure of alignment between two dendrograms. The entanglement score ranges from 0 (exact alignment) to 1 (no alignment)[29].

Defining transcription factor networks with random forests. We used random forests (a classifier that combines many single decision trees) to calculate the importance of each gene for defining cell types[30]. The varSelRF R package version 0.7–8 uses the out-of-bag error as the minimization criterion and carries out variable elimination with random forests by successively eliminating the least important variables (with importance as returned from the random forest analysis). The algorithm iteratively fits random forests, at each iteration building a new forest after discarding those variables (genes) with the smallest variable importance; the selected set of genes is the one that yields the smallest out-of-bag error rate. This leads to the selection of small sets of non-redundant variables.

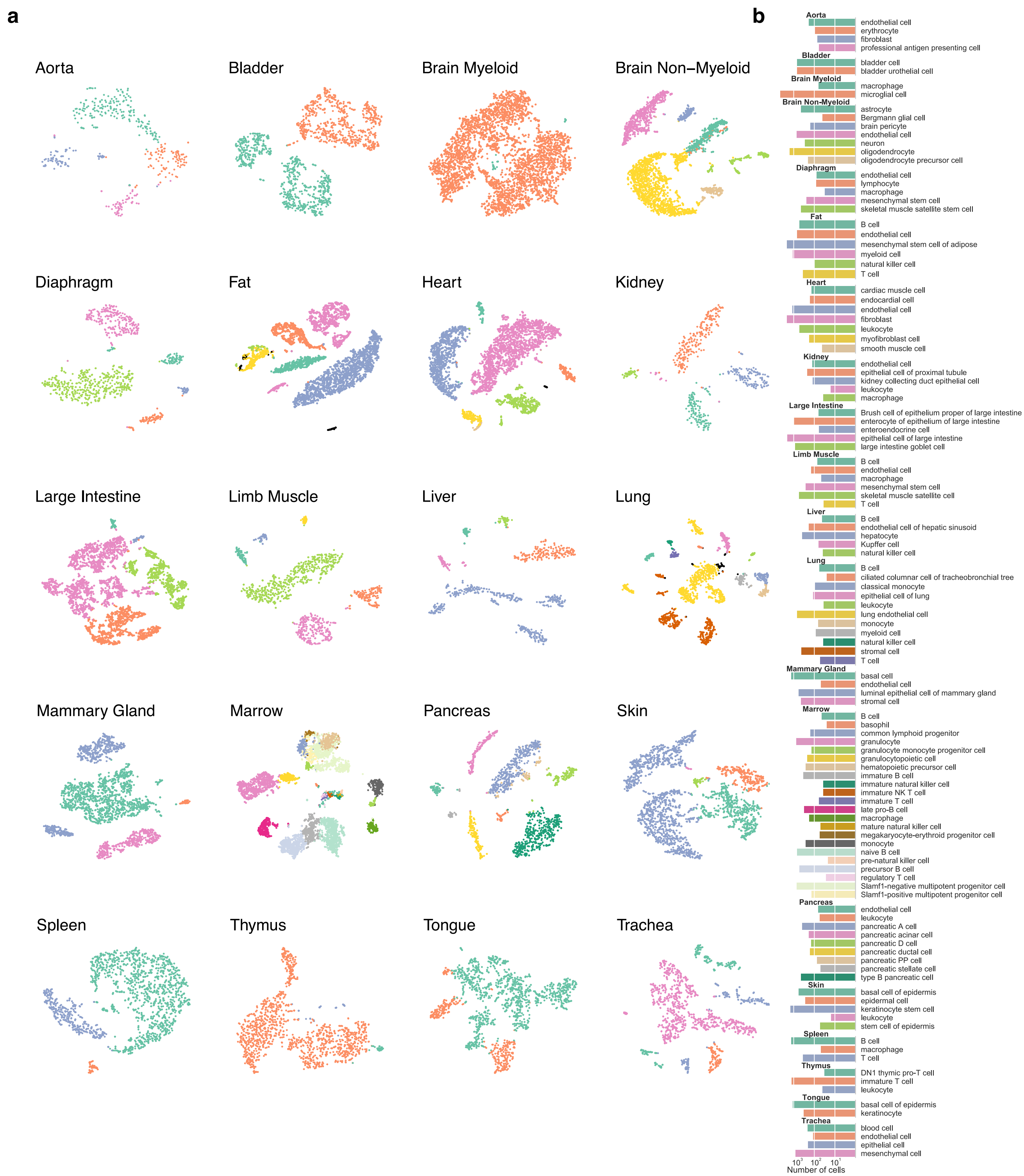
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. All code used for analysis is available on GitHub (https://github.com/czbiohub/tabula-muris).

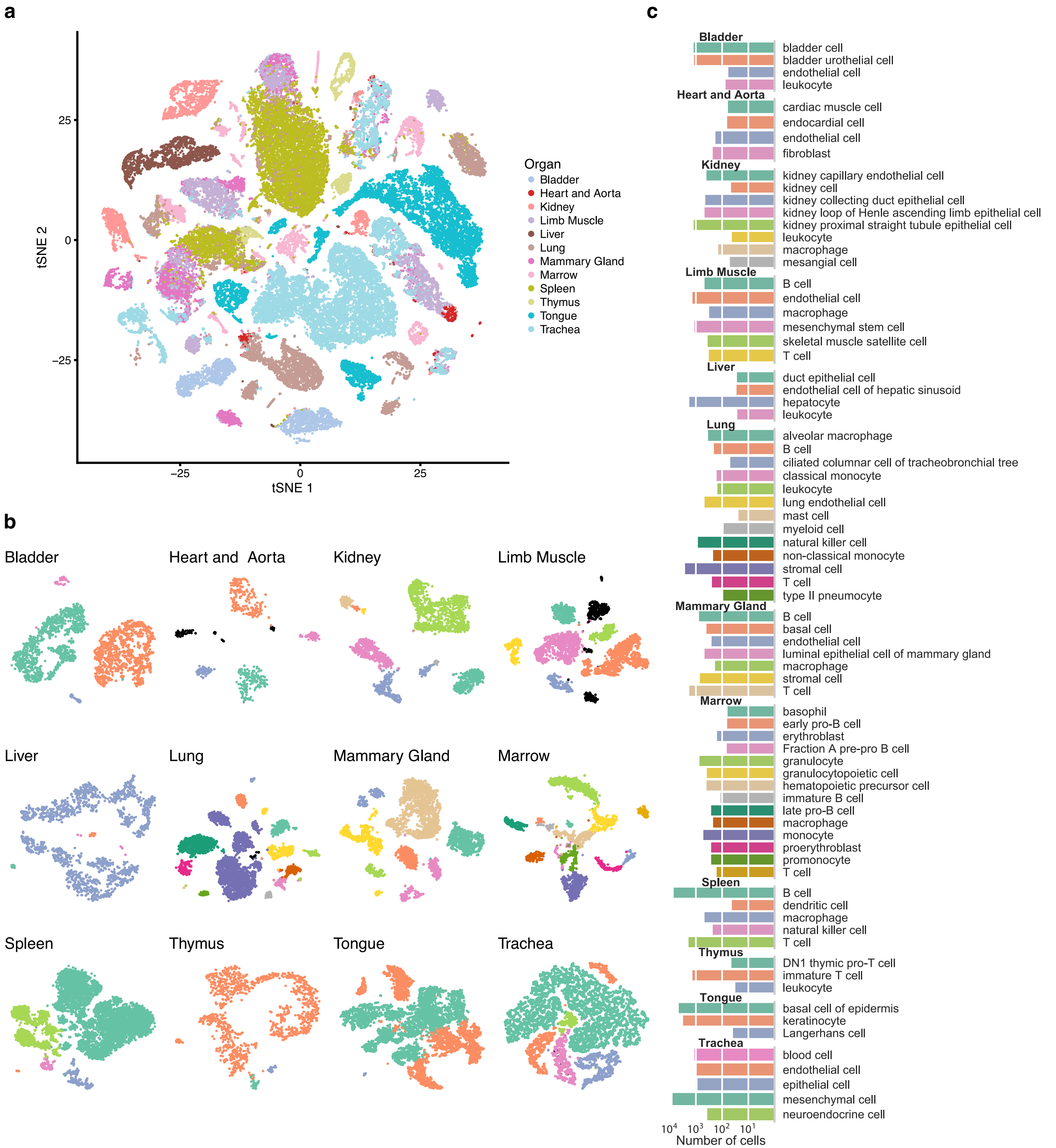
Data availability

All data, protocols and analysis scripts from the *Tabula Muris* are shared as a public resource (http://tabula-muris.ds.czbiohub.org/). Gene counts and metadata for FACS (https://doi.org/10.6084/m9.figshare.5829687.v7) and microfluidic droplets (https://doi.org/10.6084/m9.figshare.5968960.v2) from all single cells along with all produced R objects (https://doi.org/10.6084/m9.figshare.5821263.v1), as well as FACS Index data (https://doi.org/10.6084/m9.figshare.5975392) are accessible on Figshare (https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733), and raw data are available from the Gene Expression Omnibus (GSE109774).

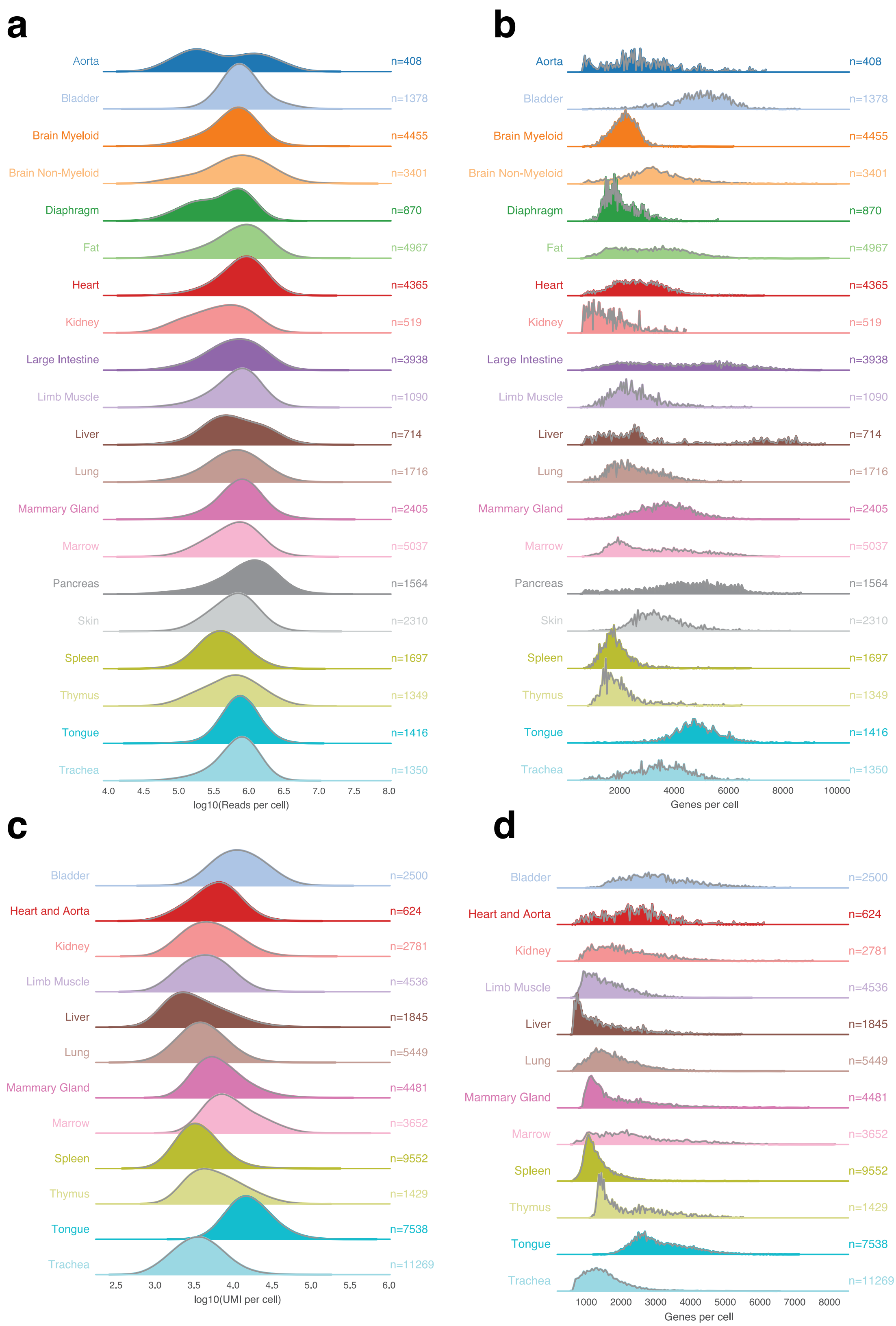
29. Kassambara, A. *Practical guide to cluster analysis in R: unsupervised machine learning* 1st edn (CreateSpace, North Charleston, 2017).
30. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3 (2006).



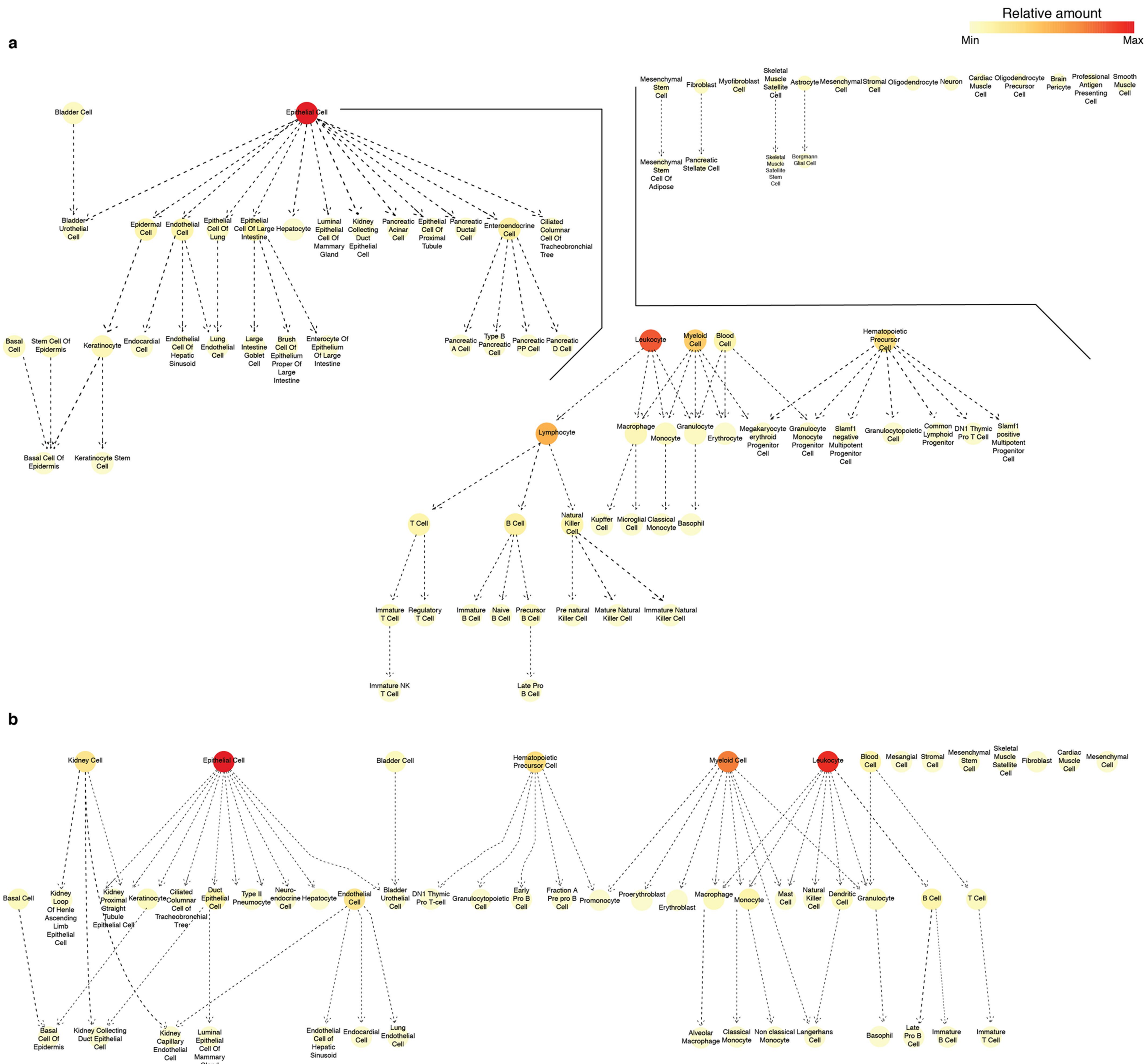
Extended Data Fig. 1 | The number and type of FACS cells that compose each organ. a, Cells for each organ visualized with *t*-SNE, coloured by cell type. Cell types were determined by differential gene expression of known markers between clusters. **b**, Bar plots quantifying the number of each annotated cell type. Cell type colours match their respective *t*-SNE plot.



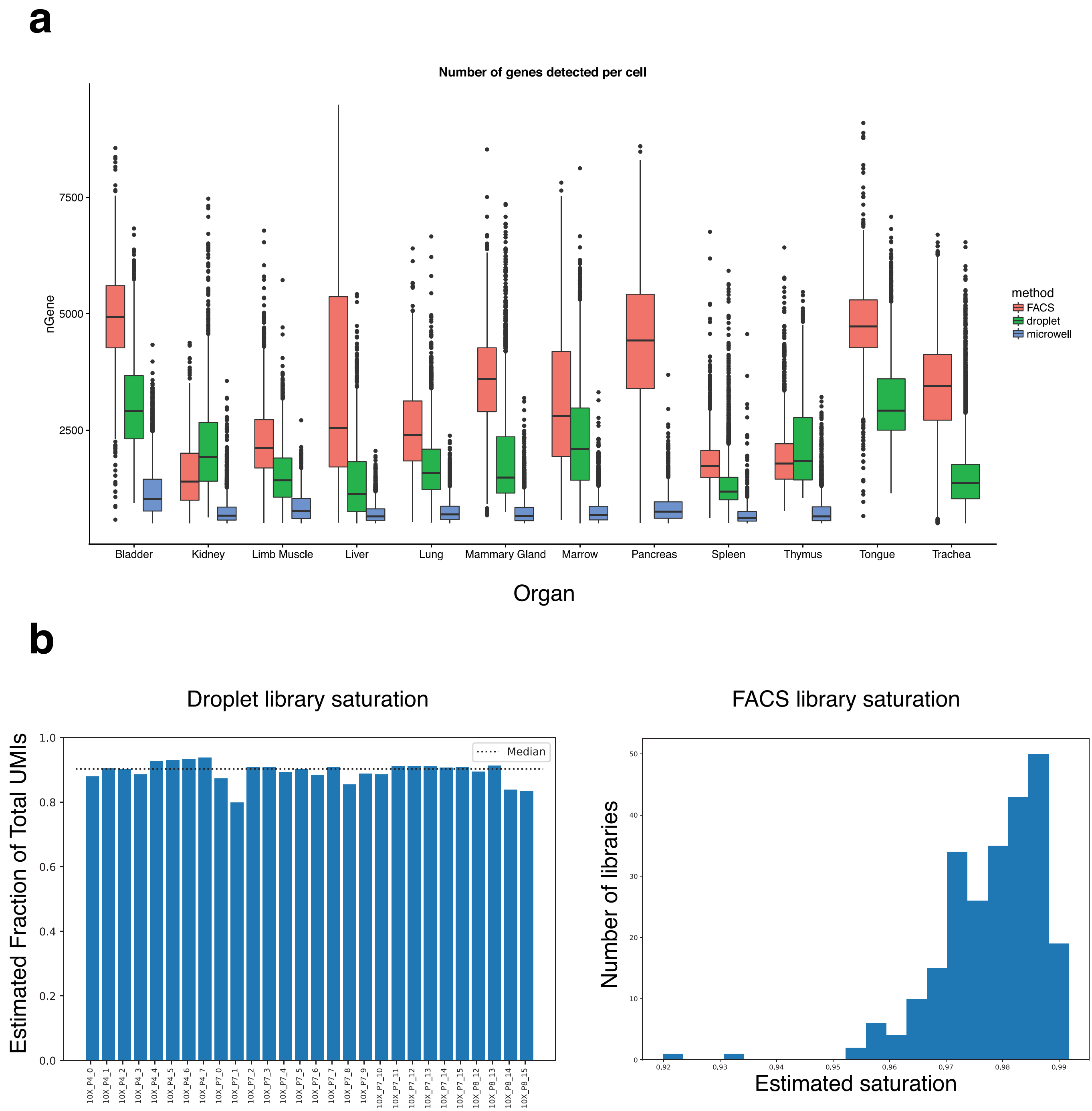
Extended Data Fig. 2 | The number and type of microfluidic cells that compose each organ. a, *t*-SNE plot of all cells collected by the microfluidic-droplet method, coloured by organ, overlaid with the predominant cell type that composes each cluster. **b**, Cells for each organ visualized with *t*-SNE, coloured by cell type. Cell types were determined by differential gene expression of known markers between clusters. **c**, Bar plots quantifying the number of each annotated cell type. Cell type colours match their respective *t*-SNE plot.



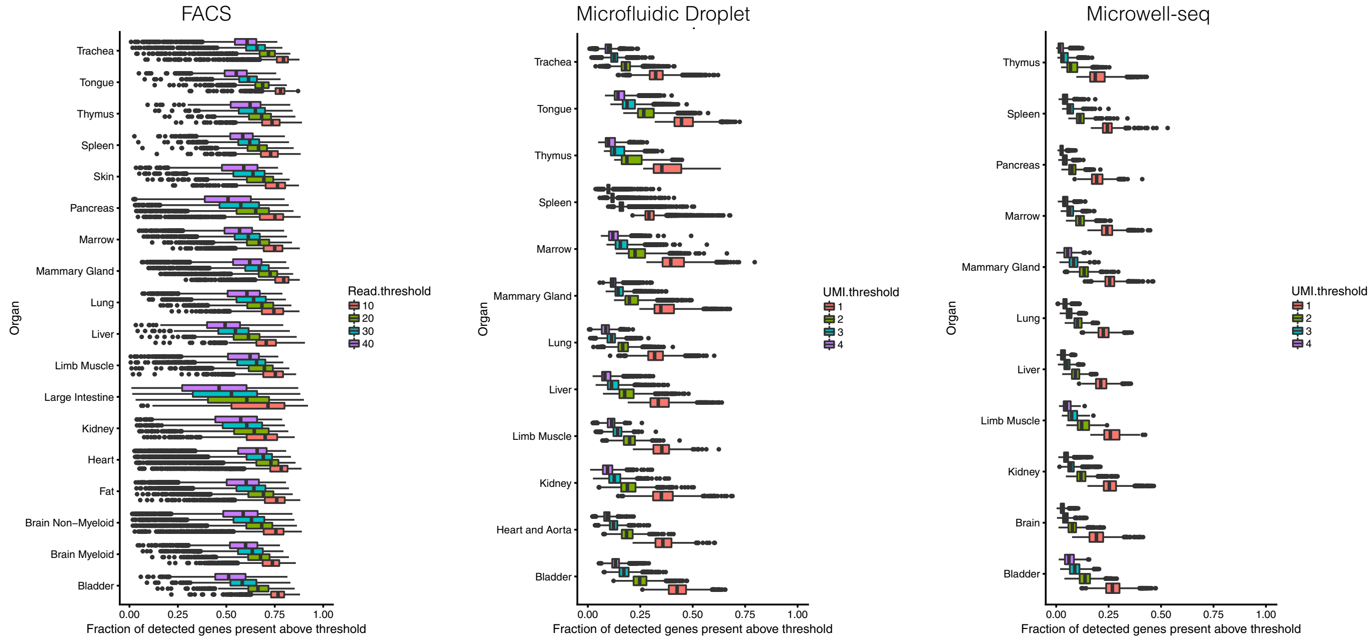
Extended Data Fig. 3 | The number of reads, UMIs and genes detected per cell for each organ. **a**, **c**, Histograms for each organ of the number of reads per cell (FACS) (**a**) and UMIs per cell (microfluidic droplet) (**c**). **b**, **d**, Histogram of the number of genes detected per cell for each organ from the FACS method (**b**), and the microfluidic-droplet method (**d**).



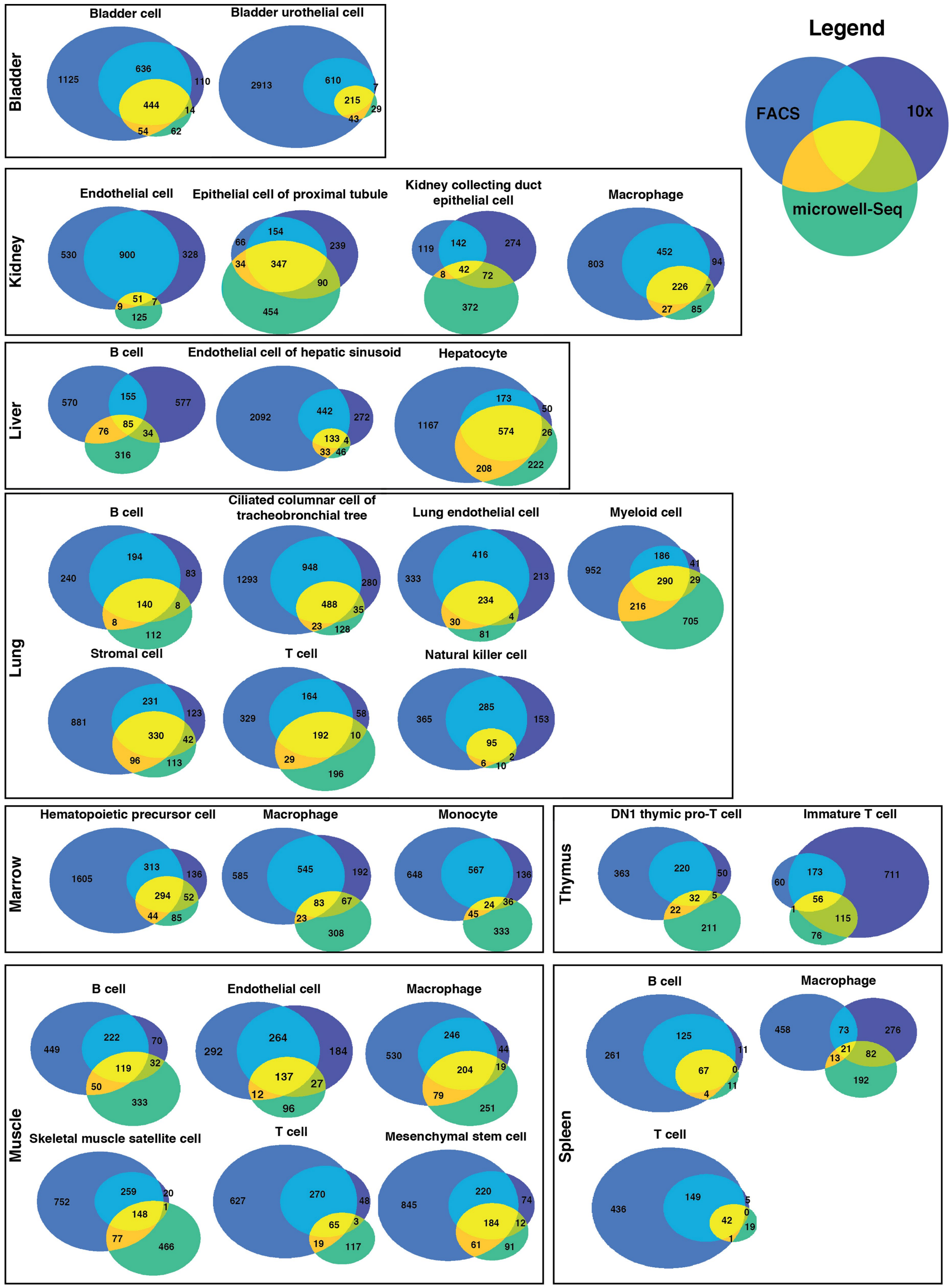
Extended Data Fig. 4 | Graphical representation of cell ontology class representation. a, b, Datasets from the FACS method (**a**) and the microfluidic-droplet method (**b**), coloured by the relative amount of each cell type in each dataset.



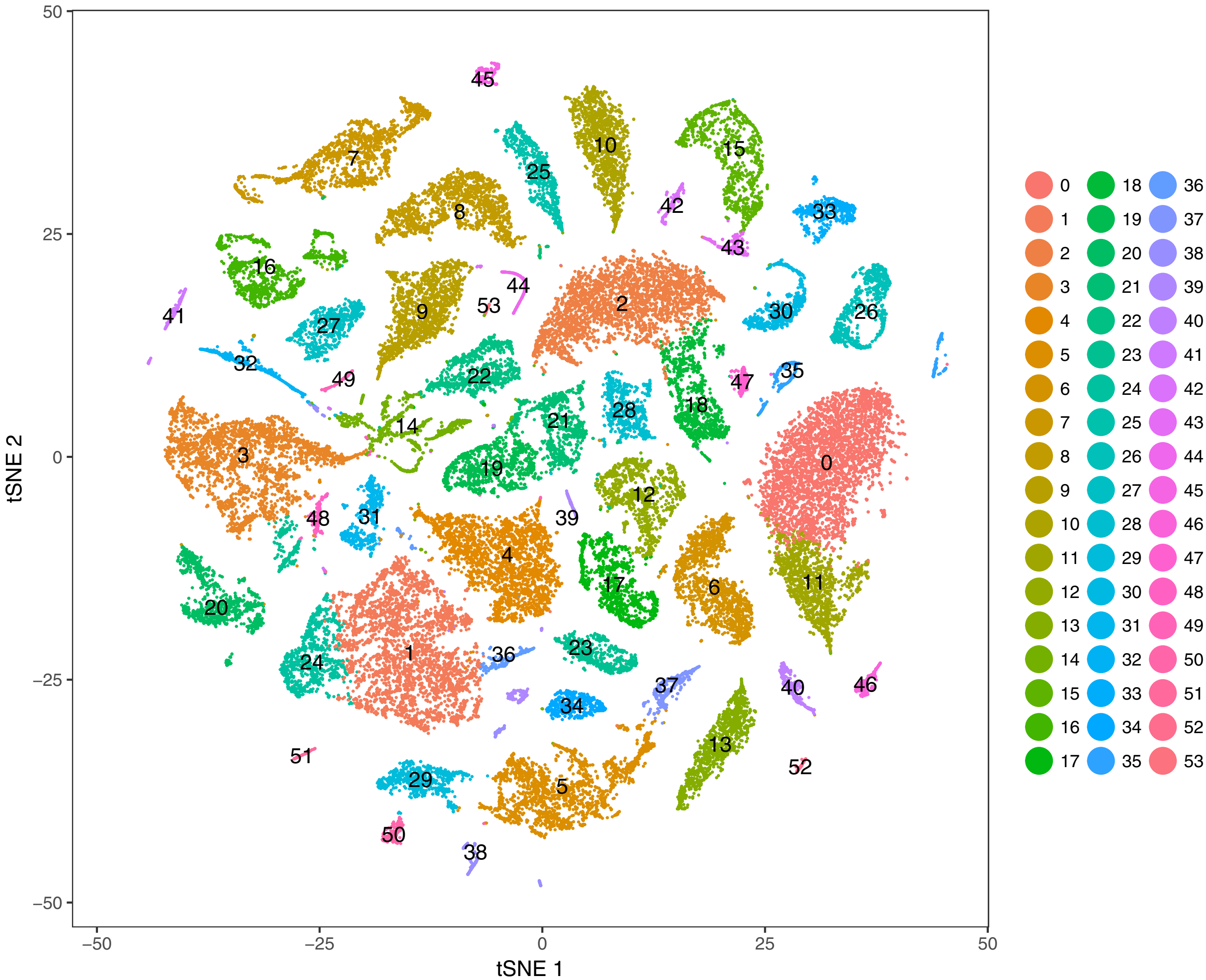
Extended Data Fig. 5 | Methodological comparison of detected genes and library saturation. a, The number of genes detected (threshold of >0 reads or UMIs per cell) by FACS (red; $n = 21,105$ individual cells), microfluidic-droplet (green; $n = 55,032$ individual cells) and microwell-seq (blue; $n = 25,891$ individual cells) methods[20]. **b**, Library saturation fraction for all microfluidic-droplet libraries. Dotted horizontal line demarcates the median saturation (around 0.9). **c**, Library saturation for all FACS libraries. Saturation was calculated using the number of detected genes while downsampling the number of reads per library. Summary statistics are contained in Supplementary Table 6.



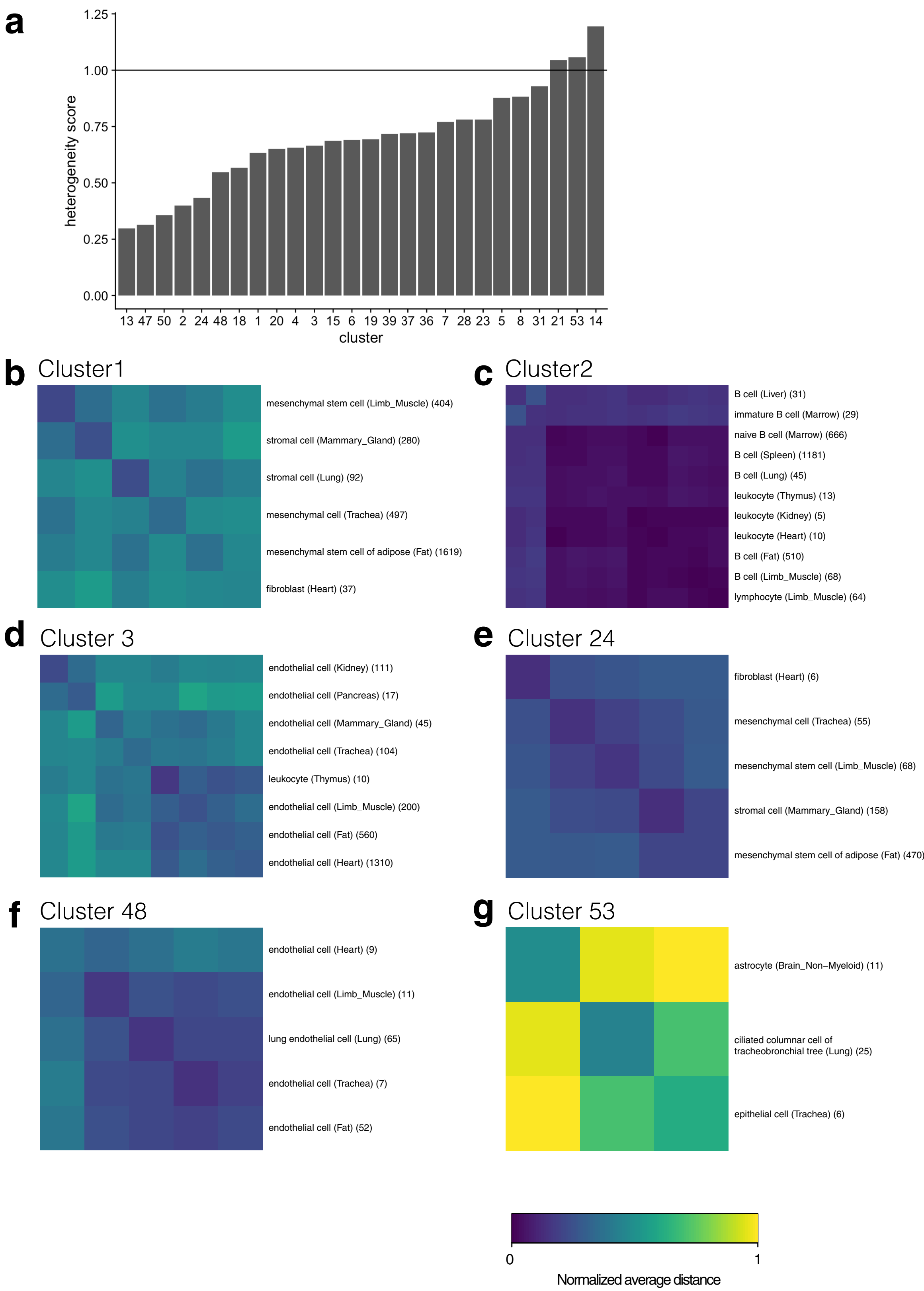
Extended Data Fig. 6 | The number of detected genes decreases similarly across organs as the read or UMI threshold is increased. Fraction of all detected genes (defined as >0 reads or UMIs) for each cell, across all organs, detected at increasing read or UMI thresholds for FACS (left; $n = 44{,}949$ individual cells), microfluidic-droplet (middle; $n = 55{,}656$ individual cells), and microwell-seq (right; $n = 28{,}372$ individual cells) methods. Summary statistics are contained in Supplementary Table 6.



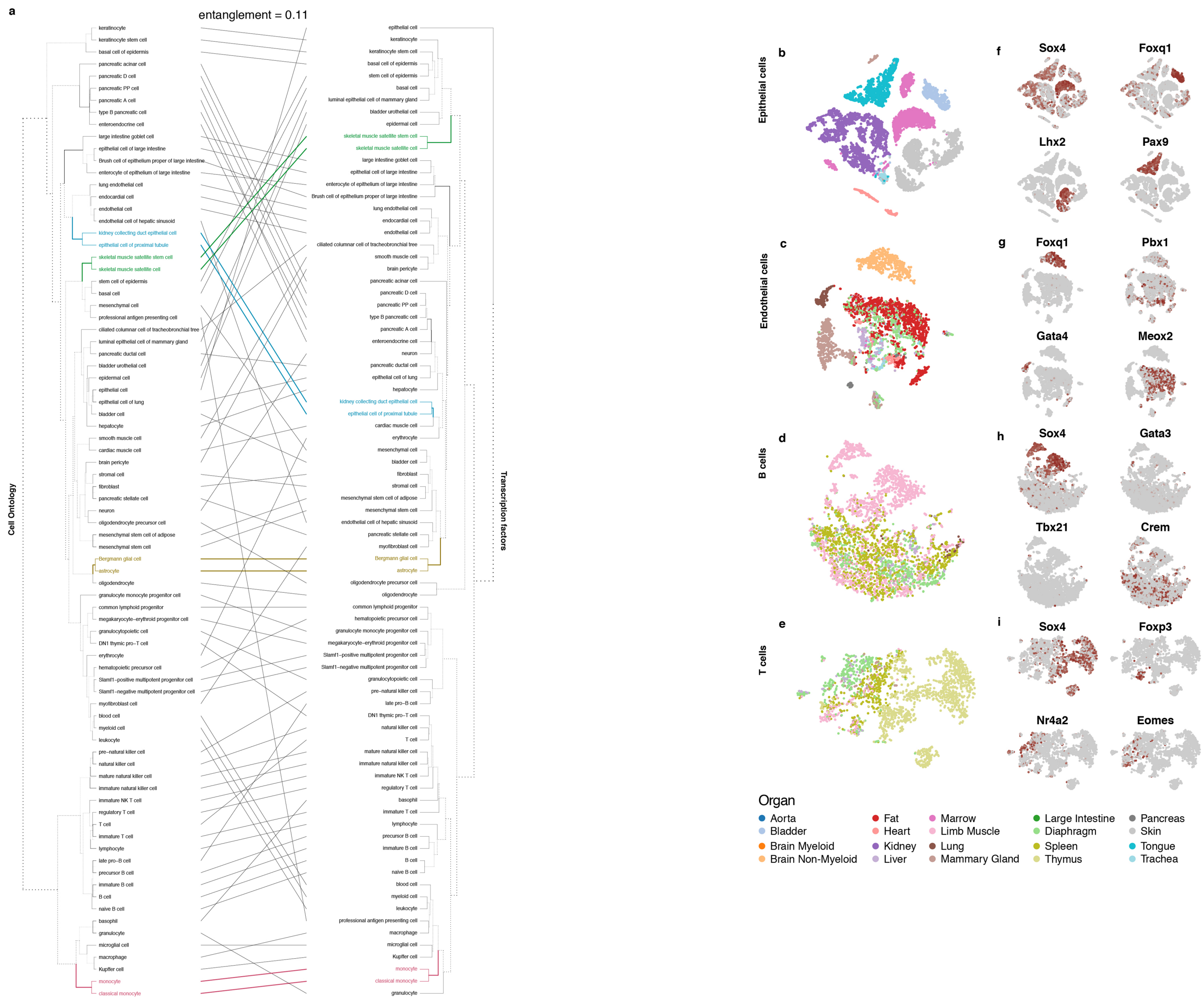
Extended Data Fig. 7 | The number of differentially expressed genes for each cell type that are common between methods. Venn diagrams showing the overlap between differentially expressed genes for each common cell type across the three methods (FACS, microfluidic-droplet and microwell-seq). Plotted data are provided in tabular form in Supplementary Table 2.



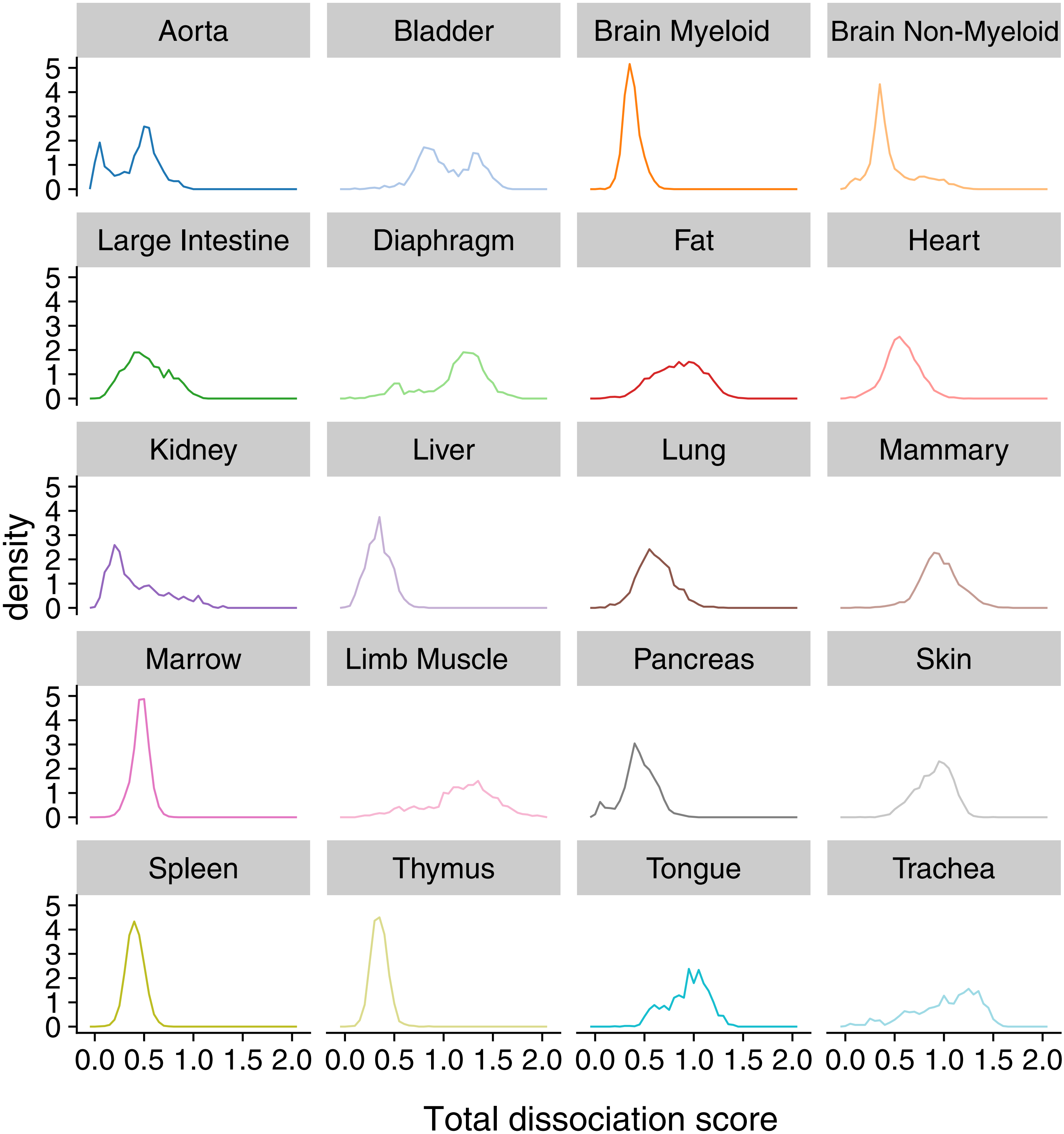
Extended Data Fig. 8 | *t*-SNE visualization of all FACS cells by cluster ID. $n = 44,949$ individual cells. Clusters are discussed in the text and further analysed in Fig. 3.



Extended Data Fig. 9 | Metrics of cluster heterogeneity. a, Bar plot showing the heterogeneity score for each cluster containing several cell types. **b–g**, Heat maps showing the average between-cell-type distances within select clusters, normalized so that the average distance between pairs of FACS cells is 1, clipped to a max of 1, for clusters 1 (**b**), 2 (**c**), 3 (**d**), 24 (**e**), 48 (**f**) and 53 (**g**).



Extended Data Fig. 10 | Contribution of transcription factors to cell identity. a, Tanglegram contrasting the dendrogram obtained using all expressed genes with one obtained using only the expression of transcription factors. The solid lines indicate segments that did not change position during the alignment between the two trees, and the dotted lines correspond to dendrogram branches reordered during the entanglement calculations. The colours indicate the branches for which identical leaves are aligned in both dendrograms. **b**–**e**, *t*-SNE visualization of epithelial (**b**), endothelial (**c**), B cells (**d**) and T cells (**e**), coloured by organ. **f**–**i**, *t*-SNE visualization of epithelial (**f**), endothelial (**g**) B cell (**h**) and T cell (**i**) expression of select transcription factors (from grey, low, to red, high). In **b**–**i**, $n = 60$ randomly selected cells for each cell type.



Extended Data Fig. 11 | Dissociation-induced gene-expression scores for each organ analysed with FACS. The dissociation score for each organ represents the magnitude of the first principal component of the 140 dissociation-associated genes from ref. [24]. The *y* axis shows the probability density of the normalized histogram.

https://doi.org/10.1038/s41586-018-0436-0

The genetic basis and cell of origin of mixed phenotype acute leukaemia

Thomas B. Alexander[1,2,37], Zhaohui Gu[3,37], Ilaria Iacobucci[3,37], Kirsten Dickerson[3], John K. Choi[3], Beisi Xu[4], Debbie Payne-Turner[3], Hiroki Yoshihara[3], Mignon L. Loh[5], John Horan[6], Barbara Buldini[7], Giuseppe Basso[7], Sarah Elitzur[8], Valerie de Haas[9], C. Michel Zwaan[9,10], Allen Yeoh[11], Dirk Reinhardt[12], Daisuke Tomizawa[13], Nobutaka Kiyokawa[14], Tim Lammens[15], Barbara De Moerloose[15], Daniel Catchpoole[16], Hiroki Hori[17], Anthony Moorman[18], Andrew S. Moore[19], Ondrej Hrusak[20], Soheil Meshinchi[21,22], Etan Orgel[23], Meenakshi Devidas[24], Michael Borowitz[25], Brent Wood[26], Nyla A. Heerema[27], Andrew Carrol[28], Yung-Li Yang[29], Malcolm A. Smith[30], Tanja M. Davidsen[31], Leandro C. Hermida[32], Patee Gesuwan[32], Marco A. Marra[33], Yussanne Ma[33], Andrew J. Mungall[33], Richard A. Moore[33], Steven J. M. Jones[33], Marcus Valentine[34], Laura J. Janke[3], Jeffrey E. Rubnitz[1], Ching-Hon Pui[1], Liang Ding[4], Yu Liu[4], Jinghui Zhang[4], Kim E. Nichols[1], James R. Downing[3], Xueyuan Cao[35], Lei Shi[35], Stanley Pounds[35], Scott Newman[4], Deqing Pei[35], Jaime M. Guidry Auvil[32], Daniela S. Gerhard[32], Stephen P. Hunger[36], Hiroto Inaba[1]* & Charles G. Mullighan[3]*

Mixed phenotype acute leukaemia (MPAL) is a high-risk subtype of leukaemia with myeloid and lymphoid features, limited genetic characterization, and a lack of consensus regarding appropriate therapy. Here we show that the two principal subtypes of MPAL, T/myeloid (T/M) and B/myeloid (B/M), are genetically distinct. Rearrangement of *ZNF384* is common in B/M MPAL, and biallelic *WT1* alterations are common in T/M MPAL, which shares genomic features with early T-cell precursor acute lymphoblastic leukaemia. We show that the intratumoral immunophenotypic heterogeneity characteristic of MPAL is independent of somatic genetic variation, that founding lesions arise in primitive haematopoietic progenitors, and that individual phenotypic subpopulations can reconstitute the immunophenotypic diversity in vivo. These findings indicate that the cell of origin and founding lesions, rather than an accumulation of distinct genomic alterations, prime tumour cells for lineage promiscuity. Moreover, these findings position MPAL in the spectrum of immature leukaemias and provide a genetically informed framework for future clinical trials of potential treatments for MPAL.

Acute leukaemia of ambiguous lineage (ALAL) comprises a collection of high-risk leukaemias defined by immunophenotype, including MPAL and acute undifferentiated leukaemia (AUL). MPAL demonstrates features of acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML), while AUL lacks lineage-defining features. MPAL represents 2–3% of cases of childhood acute leukaemia, whereas AUL is rare[1,2]. Survival rates for children and adults with MPAL are 47–75% and 20–40%, respectively, and there is no consensus regarding the optimal (AML- or ALL-directed) therapeutic regimen[1–3]. Up to 15% of patients with MPAL have rearrangements of *KMT2A* (also known as *MLL*; rearrangements referred to as *KMT2A*r) or a *BCR–ABL1* fusion gene, but the genetic basis of most cases of MPAL remains unknown. As the lineage 'aberrancy' or 'promiscuity' of T/M MPAL shares features with early T-cell precursor (ETP) ALL[4,5], we sought to define the genetic basis of MPAL, to compare its genomic landscape to those of other leukaemia subtypes, and to determine the genetic basis of the intratumoral phenotypic heterogeneity that is characteristic of this disorder.

Genomic characterization of ALAL

We performed a central review of 159 potential paediatric cases of ALAL by repeating ($n = 138$) or reviewing flow cytometry data ($n = 21$); 115 fulfilled WHO (World Health Organization) criteria for the diagnosis of ALAL[6] (Extended Data Fig. 1). There was a male pre-

[1]Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN, USA. [2]Department of Pediatrics, University of North Carolina, Chapel Hill, NC, USA. [3]Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN, USA. [4]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, USA. [5]Department of Pediatrics, Benioff Children's Hospital and the Helen Diller Family Comprehensive Cancer Center, University of California at San Francisco, San Francisco, CA, USA. [6]Aflac Cancer and Blood Disorders Center, Children's Healthcare of Atlanta and Emory University School of Medicine, Department of Pediatrics, Atlanta, GA, USA. [7]Department of Women and Child Health, Hemato-Oncology Division, University of Padova, Padova, Italy. [8]Pediatric Hematology-Oncology, Schneider Children's Medical Center, Sackler Faculty of Medicine, Tel Aviv University, Israel. [9]Prinses Maxima Centre, Utrecht, The Netherlands. [10]Department of Pediatric Oncology, Erasmus MC-Sophia, Rotterdam, The Netherlands. [11]Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [12]Universitäts-Klinikum, Essen, Germany. [13]Division of Leukemia and Lymphoma, Children's Cancer Center, National Center for Child Health and Development, Tokyo, Japan. [14]Department of Pediatric Hematology and Oncology Research, National Research Institute for Child Health and Development, Tokyo, Japan. [15]Department of Pediatric Hematology-Oncology and Stem Cell Transplantation, Ghent University Hospital, Ghent, Belgium. [16]The Tumour Bank CCRU, The Kids Research Institute, The Children's Hospital at Westmead, Westmead, New South Wales, Australia. [17]Department of Pediatrics, Mie University, Tsu, Japan. [18]Wolfson Childhood Cancer Centre, Northern Institute for Cancer Research, Newcastle University, Newcastle-upon-Tyne, UK. [19]The University of Queensland Diamantina Institute & Children's Health, Brisbane, Queensland, Australia. [20]Department of Paediatric Haematology and Oncology, 2nd Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic. [21]Fred Hutchinson Cancer Research Center, Clinical Research Division, Seattle, WA, USA. [22]Children's Oncology Group, Arcadia, CA, USA. [23]Children's Center for Cancer and Blood Disease, Children's Hospital Los Angeles, Los Angeles, CA, USA. [24]University of Florida, Gainesville, FL, USA. [25]Johns Hopkins Medical Institutions, Baltimore, MD, USA. [26]University of Washington, Seattle, WA, USA. [27]The Ohio State University School of Medicine, Columbus, OH, USA. [28]University of Alabama at Birmingham, Birmingham, AL, USA. [29]Department of Laboratory Medicine and Pediatrics, National Taiwan University Hospital, College of Medicine, National Taiwan University, Taipei, Taiwan. [30]Cancer Therapy Evaluation Program, National Cancer Institute, Bethesda, MD, USA. [31]Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD, USA. [32]Office of Cancer Genomics, National Cancer Institute, Bethesda, MD, USA. [33]Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. [34]Cytogenetics Shared Resource, St. Jude Children's Research Hospital, Memphis, TN, USA. [35]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA. [36]Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia and the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [37]These authors contributed equally: Thomas B. Alexander, Zhaohui Gu, Ilaria Iacobucci. *e-mail: hiroto.inaba@stjude.org; charles.mullighan@stjude.org

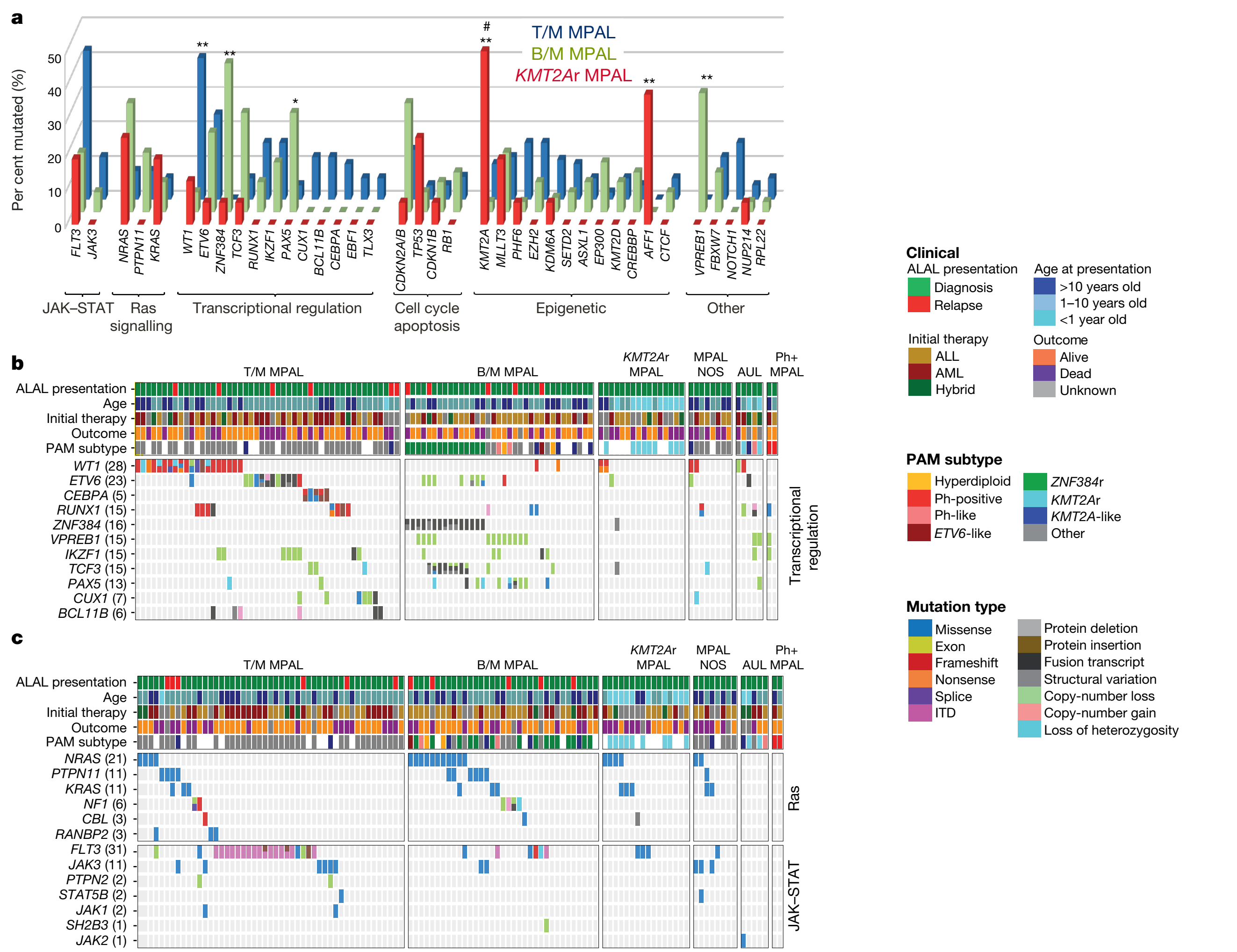


Fig. 1 | Genomic overview of ALAL. a, Distribution of the most frequently altered genes by MPAL subtype. Frequency of mutations in the different MPAL subtypes were compared by two-sided Fisher exact tests; $^{**}P < 0.001$, $^{*}0.001 < P < 0.01$ (see Supplementary Table 13 for numbers for each group and P values for each gene). #*KMT2A* alterations were present in all cases in the *KMT2A*r subgroup. **b**, Oncoprint of mutations in transcriptional regulation and cell cycle/apoptosis pathways. ITD, internal tandem duplication; PAM, prediction analysis of microarrays. **c**, Oncoprint of mutations in signalling pathways. Mutations altering genes involved in transcription and signalling pathways in these subtypes are distinct.

dominance of ALAL (1.6:1), which was diagnosed at similar frequency throughout childhood, except for cases with *KMT2A*r, which were common in infants (Supplementary Tables 1, 2). The cohort included 49 cases of T/M MPAL, 35 B/M MPAL, 16 *KMT2A*r MPAL and 2 *BCR–ABL1* MPAL, 8 MPAL not otherwise specified (NOS), and 5 AUL. There was extensive immunophenotypic heterogeneity, with bilineal patterns (multiple immunophenotypic subpopulations), biphenotypic patterns (coexpression of lymphoid and myeloid antigens), or both (Extended Data Fig. 2a–g). There was no difference in five-year overall survival between T/M MPAL and B/M MPAL (56.7% ± 10.8% (95% confidence interval) and 59.7% ± 11.4%, respectively); outcome for patients with *KMT2A*r was poor (five-year overall survival 21.2% ± 10.8%) (Extended Data Fig. 2h–o).

Genomic alterations were examined by exome ($n = 92$), transcriptome ($n = 95$), and/or whole-genome ($n = 47$) sequencing, and single nucleotide polymorphism (SNP) array analysis ($n = 95$) (Supplementary Tables 3, 4). We identified 158 recurrently altered genes, of which 81 were mutated in at least three cases. Commonly mutated genes included those recurrent in AML, such as *FLT3* ($n = 31$), *RUNX1* ($n = 15$), *CUX1* ($n = 7$) and *CEBPA* ($n = 5$); those recurrent in ALL, including *CDKN2A* or *CDKN2B* ($n = 22$), *ETV6* ($n = 23$), and *VPREB1* ($n = 15$); and those recurrent in both AML and ALL, including *WT1* ($n = 28$) and *KMT2A* ($n = 26$) (Fig. 1a, Extended Data Figs. 3, 4 and Supplementary Tables 5–13). We analysed associations between genomic alterations and age at diagnosis, sex and disease subtype, and between pathway alterations and outcome (Supplementary Tables 14, 15 and Supplementary Note). We analysed germline samples for potential pathogenic variants in recurrently somatically mutated genes, and identified few putatively deleterious variants[7] (Supplementary Table 16 and Supplementary Note).

Distinct profiles of MPAL subtypes

The three most common subtypes of MPAL (T/M, B/M and *KMT2A*r) had distinct patterns of genomic alterations (Fig. 1a–c and Supplementary Table 13). As in infant ALL, *KMT2A*r MPAL had a low mutation burden (median 1 (range 0–3) copy number alterations (CNAs) and 4 (0–12) single nucleotide variants (SNVs) or insertions/deletions (indels) per case), whereas the mutation burden was higher for T/M MPAL (4.5 (0–35) CNAs, 8 (2–29) SNVs or indels) and B/M MPAL (3.5 (0–29) CNAs, 9 (0–167) SNVs or indels) (Extended Data Fig. 3b). Alterations in genes encoding transcriptional regulators were detected in 100% of cases of T/M MPAL, with mutually exclusive alterations in *WT1*, *ETV6*, *RUNX1* and *CEBPA* in 82% of cases (Fig. 1b and Extended Data Fig. 5a, b); and in 94% of cases of B/M MPAL, with the B-lineage transcriptional regulators *PAX5* and *IKZF1* altered in 40% of cases (Fig. 1b).

Alterations in signalling pathways were observed in 88% of cases of T/M MPAL, 74% of cases of B/M MPAL and 63% of cases of *KMT2A*r MPAL. Alterations in JAK–STAT signalling were more common in

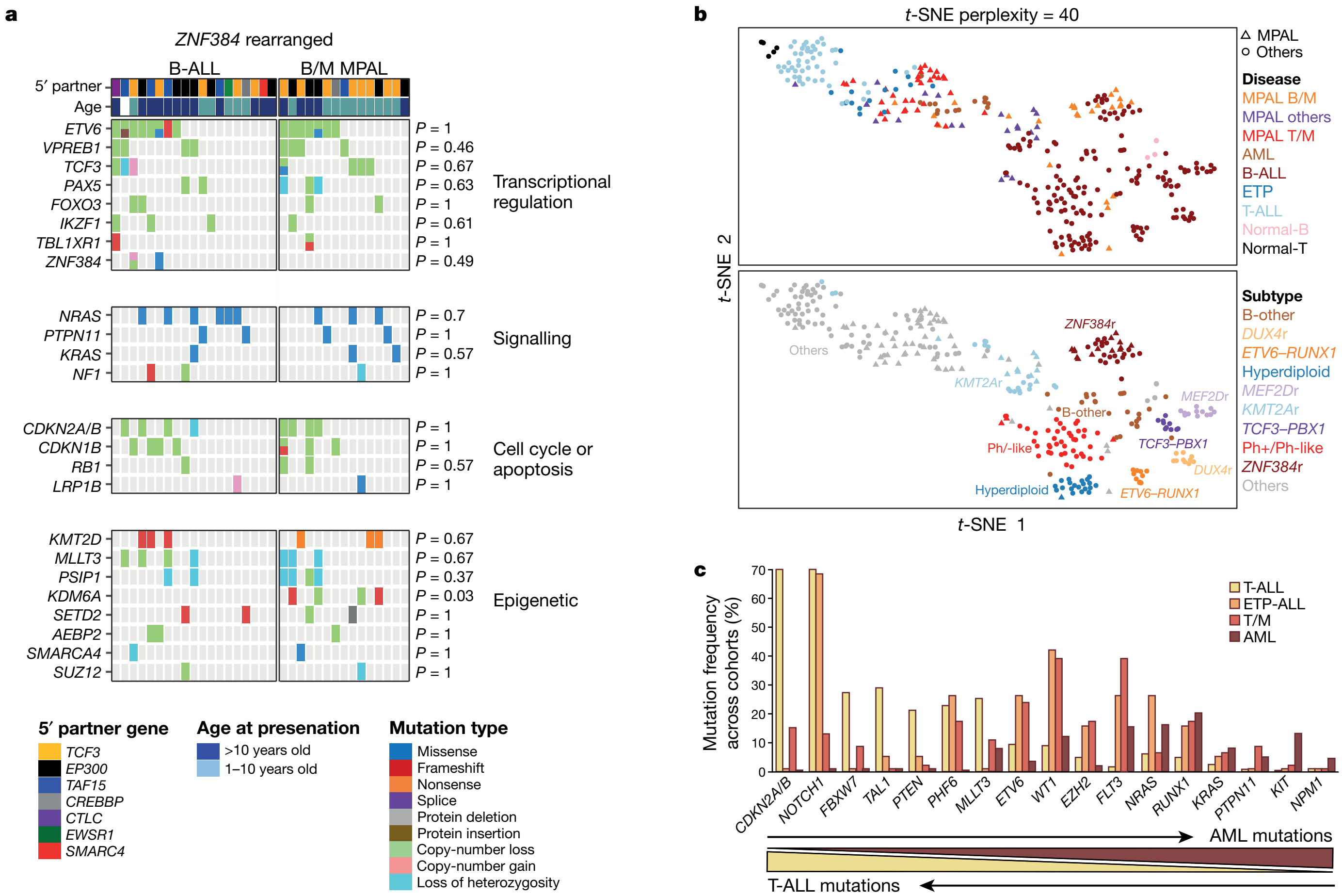


Fig. 2 | Genomic comparisons across leukaemia subtypes. a, Mutations observed in *ZNF384*r B-ALL ($n = 19$) and *ZNF384*r B/M MPAL ($n = 15$), showing similar mutational profile between the two phenotypically defined subtypes. **b**, *t*-distributed stochastic neighbour embedding (*t*-SNE) plot of top 1,000 variably expressed genes of ALAL, B-ALL, T-ALL, ETP-ALL, AML, and normal lymphocytes, showing that B/M MPAL has a GEP more similar to B-ALL than AML, and T/M MPAL more similar to ETP-ALL than AML. *ZNF384*r cases cluster together, without separation based upon B/M MPAL or B-ALL phenotype. Cases in the ALAL-other category, including *KMT2A*r MPAL, AUL, and MPAL NOS, are intermixed across the transcriptional continuum, primarily between AML and B-ALL clusters. **c**, Depiction of the frequency of mutations of the five most frequently altered genes from each disease cohort, demonstrating that T/M MPAL ($n = 49$) and ETP-ALL ($n = 19$) share a high frequency of mutations in *ETV6*, *WT1*, *EZH2* and *FLT3*, while lacking the most characteristically mutated genes in T-ALL ($n = 245$) and AML ($n = 197$).

T/M MPAL (57%) than B/M MPAL (23%) or *KMT2A*r MPAL (19%) (Fig. 1c), and we observed a negative correlation between alterations in *FLT3* (43%) and the Ras pathway (33%) in T/M MPAL ($P = 0.002$) (Fig. 1c and Supplementary Table 15). Ras pathway alterations were common in B/M MPAL (63%, most commonly *NRAS* and *PTPN11*). Genes encoding epigenetic regulators were mutated in 69% of cases of T/M MPAL, including inactivating mutations in *EZH2*[5] (16%) and *PHF6* (16%), and in 63% of cases of B/M MPAL, most commonly in *MLLT3* (17%), *KDM6A* (in one-third of *ZNF384*-rearranged cases), *EP300* and *CREBBP* (Supplementary Table 13).

Transcriptome sequencing identified chimaeric in-frame fusions in 15 of 40 cases of T/M MPAL: *ZEB2–BCL11B* ($n = 3$), *ETV6–NCOA2* ($n = 2$), *ETV6–ARNT* ($n = 2$) and single cases of *ETV6–FOXO1*, *ETV6–MAML3*, *NUP214–ABL1*, *PICALM–MLLT10* and *PCM1–FGFR1* (Supplementary Tables 17–20). *KMT2A*r MPAL had a B/M phenotype in 15 out of 16 cases and a T/M phenotype in one case, and involved *AFF1* (also known as *AF4*) in seven cases, *MLLT3* (also known as *AF9*) in three cases and *MLLT1* (also known as *ENL*) in two cases. *KMT2A*r was also found in two of five cases of AUL.

ZNF384 rearrangement in leukaemia

Rearrangement of *ZNF384 (ZNF384*r*)* was present in 48% of cases of B/M MPAL, involving *TCF3* ($n = 8$), *EP300* ($n = 5$), *TAF15* ($n = 1$) and *CREBBP* ($n = 1$) (Extended Data Fig. 5c). The chimaeric fusions involved the entire ZNF384 coding region, loss of the C termini of the partner genes, and translation of both wild-type ZNF384 and chimaeric fusion proteins. The mutational burden of *ZNF384*r B/M MPAL (median of 4 (1–29) CNAs and 8 (3–39) SNVs or indels) was similar to those of other MPAL subtypes (Extended Data Fig. 3c), with no variation in mutations between immunophenotypic subpopulations in ten cases examined (Extended Data Fig. 5d). *ZNF384*r, most commonly with *TCF3*, is also observed in B cell ALL (B-ALL), in which aberrant expression of myeloid markers that do not fulfil the diagnostic criteria for B/M MPAL is common[8]. The genomic landscape of childhood *ZNF384*r B-ALL ($n = 19$) (Supplementary Tables 21, 22) was similar to that of *ZNF384*r MPAL with the exception of *KDM6A* alterations, which were observed only in *ZNF384*r MPAL (Fig. 2a). Analysis of a diverse range of acute leukaemias, including AML (Supplementary Tables 23, 24), showed that the gene expression profiles (GEPs) of *ZNF384*r B/M MPAL and B-ALL were indistinguishable (Fig. 2b, Extended Data Fig. 5e and Supplementary Table 25). Patients with *ZNF384*r exhibited higher *FLT3* expression than those with other types of B/M or T/M MPAL (Extended Data Fig. 5f). Cases of B/M MPAL that exhibited genomic features of other subtypes of B-ALL, such as hyperdiploidy or a Ph-like GEP, clustered with those subtypes of B-ALL (Fig. 2b). Gene set enrichment analysis suggested that *ZNF384*r B/M MPAL was arrested at a more mature stage of development than other types of B/M MPAL (Extended Data Fig. 6a and Supplementary

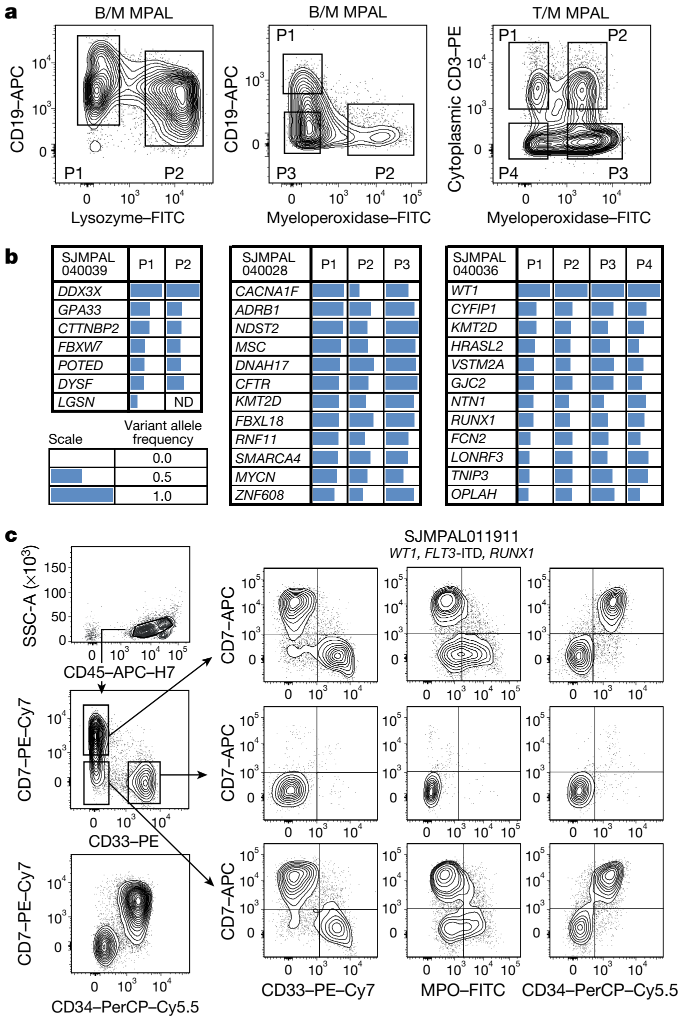


Fig. 3 | Plasticity is independent of mutation. a, Flow cytometry scatter plots of representative cases of MPAL, showing the primary lymphoid marker (CD19 or cytoplasmic CD3) and myeloid marker (myeloperoxidase or lysozyme) used for sorting subpopulations. P1–P4 represent sorted subpopulations subjected to DNA sequencing. **b**, Variant allele frequency (VAF, represented by length of blue bar) from each of the purified populations in **a** demonstrating concordance of mutational VAF of SNV or indel between distinct immunophenotypically defined subpopulations. ND, not detected. **c**, Phenotypic subpopulations from case SJMPAL011911 were sorted (first column) and injected into irradiated NSG-SGM3 mice. Key gene alterations are show above the flow scatter plots. Remaining plots show the immunophenotype of harvested bone marrow of engrafted leukaemia from each starting subpopulation, demonstrating recapitulation of mixed phenotype leukaemia from two sorted subpopulations. The third subpopulation (CD7$^-$, CD33$^+$) also engrafted with hCD45$^+$ cells and morphologic leukaemia, but with an undifferentiated immunophenotype. SSC-A, side scatter area.

Tables 26, 27). However, compared with B-ALL, *ZNF384*r leukaemia showed enrichment of stem cell pathways and genes upregulated in ETP-ALL (Extended Data Fig. 6b and Supplementary Tables 27–29). Serial sampling of a case of *ZNF384*r B/M MPAL showed acquisition of a focal heterozygous *IKZF1* deletion at first relapse, and a focal homozygous deletion of *CDKN2A* and *CDKN2B* at second relapse, with a shift from a myeloid to a lymphoid immunophenotype. Thus, *ZNF384*r defines a distinct subtype of acute leukaemia with a variable immunophenotype ranging from B-ALL to B/M MPAL.

To further investigate the role of *ZNF384* rearrangement in leukemogenesis, we expressed haemagglutinin (HA)-tagged ZNF384, TAF15–ZNF384 and TCF3–ZNF384 in mouse *Arf*$^{-/-}$ pre-B cells[9] (Extended Data Fig. 6c). Chromatin immunoprecipitation with sequencing (ChIP–seq) identified 2,298 peaks with new or increased binding of the fusion proteins compared to wild-type ZNF384, and 495 peaks with reduced binding (Extended Data Fig. 6d). Gained or increased peaks contained the core ZNF384 binding motif, and were enriched at promoters of genes important for immune system development and transcriptional regulation (Supplementary Tables 30, 31). Increased promoter binding was associated with increased gene expression (Extended Data Fig. 6e and Supplementary Table 32), with similarity between the GEPs of mouse pre-B cells expressing ZNF384 fusions and human *ZNF384*r leukaemia cells (Extended Data Fig. 6f and Supplementary Table 28). Thus, chimaeric ZNF384 oncoproteins exhibit perturbed binding and drive transcriptional deregulation in human *ZNF384*r leukaemia.

The driver alterations and GEPs of non-*ZNF384*r MPAL and AUL were heterogeneous (Supplementary Table 33). Three cases were Ph-like (*EBF1–PDGFRB, IGH–CRLF2* and a case lacking an identified kinase lesion), and two were hyperdiploid. Eight cases were *KMT2A*-like with HOXA9 deregulation, and six of these had genetic alterations associated with HOXA overexpression: *MLLT10* rearrangement ($n=2$), *SET–NUP214* ($n=2$), *KMT2A* partial tandem duplication and *MNX1–ETV6* ($n=1$ each).

Similarity between T/M MPAL and ETP-ALL

ETP-ALL exhibits aberrant expression of stem cell and myeloid markers (with the exception of myeloperoxidase, which would classify the disease as AML or MPAL)[10]. ETP-ALL is characterized by mutations in regulators of haematopoietic development, signalling, and chromatin remodelling, and a GEP suggesting the cell of origin to be a haematopoietic stem cell (HSC) or progenitor, rather than a T-cell precursor[5]. Because T/M MPAL and ETP-ALL are defined by a phenotype that includes lymphoid and myeloid features[6,10], we hypothesized that they might share molecular features. We compared the genomic features of T/M MPAL with those of childhood T cell ALL (T-ALL; $n=245$)[11], ETP-ALL ($n=19$)[11] and AML ($n=197$)[12] (Supplementary Tables 34, 35). Transcription factor gene alterations were common in each but varied between subtypes (Extended Data Fig. 6g). The core transcription factors driving T-ALL (*TAL1*, *TAL2*, *TLX1*, *TLX3*, *LMO1*, *LMO2*, *NKX2-1*, *HOXA10* and *LYL1*) were less frequently altered in T/M MPAL and ETP-ALL (63% versus 16% and 26%, respectively; $P<0.001$). Alterations that deregulated *TAL1*, which were present in 31% of cases of T-ALL, were never observed in T/M MPAL, including the 15 cases for which whole-genome sequencing data were examined for noncoding enhancer mutations[13]. Other alterations that are common in T-ALL, such as *MYB* amplification, *LEF1* deletion, *CDKN2A* and *CDKN2B* deletions, and amplification of the NOTCH1-driven *MYC* enhancer, were rare in T/M MPAL and ETP-ALL. By contrast, *WT1* alterations were common in T/M MPAL (41%) and ETP-ALL (42%), but not in T-ALL (9%; $P<0.001$). Alterations of *CEBPA* and *CUX1* were common in T/M MPAL but not ETP-ALL or T-ALL. Conversely, *NOTCH1* mutations were uncommon in T/M MPAL and AML. Signalling pathway mutations were also associated with specific subtypes, with Ras and JAK–STAT pathway mutations being common in T/M MPAL and ETP-ALL, and phosphotidylinositol 3-kinase (PI3K) signalling pathway mutations being common in T-ALL (Extended Data Fig. 6h). Several genes were mutated at similar frequencies in T/M MPAL and ETP-ALL, including *ETV6*, *EZH2*, *WT1* and *FLT3* (Fig. 2c), and the GEPs of T/M MPAL and ETP-ALL were similar (Fig. 2b). Thus, T/M MPAL and ETP-ALL are similar entities in the spectrum of immature leukaemias.

Analysis of intratumoral variegation

Elucidating whether the intra-sample immunophenotypic heterogeneity is determined by genetic variegation or by genomic priming of a haematopoietic progenitor has important implications for therapy. Accordingly, we sequenced 2–4 subpopulations from 50 cases of MPAL (Supplementary Table 36). In 41 cases, the non-silent mutations were present in each separate population (Fig. 3a, b and Extended Data Fig. 7a). In nine cases, multiple mutations were detected

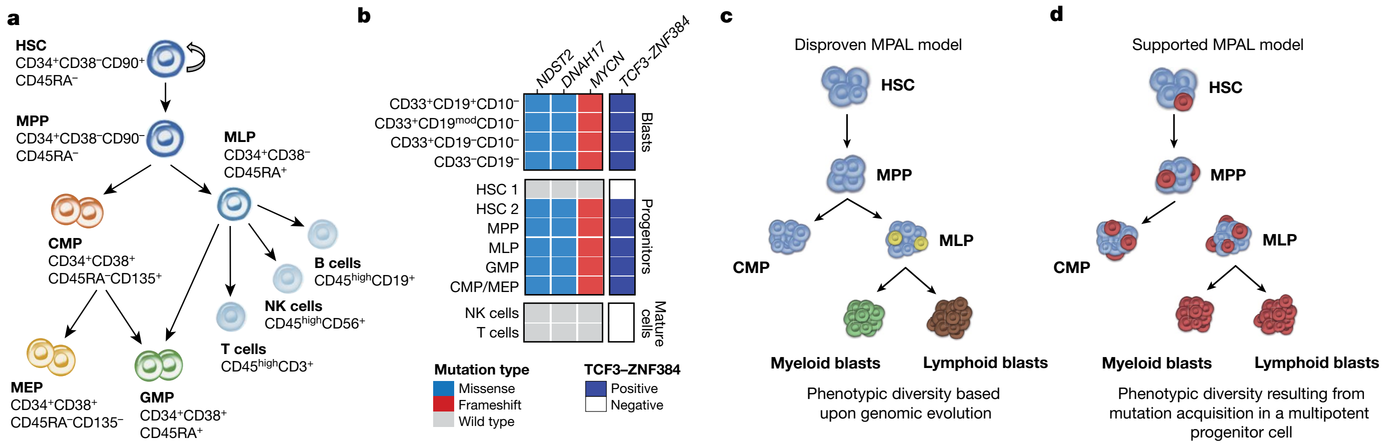


Fig. 4 | Model of MPAL leukaemogenesis. a, Schematic and simplified representation of human haematopoietic hierarchy showing HSCs, multipotent progenitors (MPPs), multilymphoid progenitors (MLPs), megakaryocyte erythroid progenitors (MEPs), common myeloid progenitors (CMPs), granulocyte monocyte progenitors (GMPs), and mature lymphocytes: B cells, T cells, and NK cells. **b**, Summary of the presence of *ZNF384*r and additional somatic alterations in isolated stem/progenitor, mature and blast cell populations showing the presence of each alteration throughout haematopoietic development. **c**, **d**, Potential models of bilineal MPAL leukaemogenesis. Different colours represent clones with different genomic alterations. **c**, A model of MPAL in which phenotypic divergence is driven by acquisition of secondary genomic alterations (yellow and green cells), which is inconsistent with the results of the current study. **d**, A model of MPAL showing that necessary and sufficient mutations are acquired in an early haematopoietic progenitor that retains myeloid and lymphoid potential, thus propagating similar mutation profiles in the different phenotypes. The results of the current study support this model of leukaemogenesis.

in a single gene (*WT1* in five cases) with at least one of the mutations detected in all subpopulations in all cases. In two cases, the second mutation called from the same gene was not present in each subpopulation sequenced (*WHSC1* in T/M case SJMPAL016447 and *CREBBP* in T/M case SJMPAL017976). In five cases, a subpopulation-restricted mutation occurred in a signalling pathway, either as gain of function (*PTPN11*, *FLT3*) or loss of function (*NF1*, *CBL*) (Supplementary Table 36), consistent with previous studies of diagnosis and relapse pairs showing frequent subclonal signalling alterations[14]. By contrast, mutations in the most commonly altered transcription factor in T/M MPAL, *WT1*, were consistently present in the major clone in each case. These observations support the notion that transcription factor gene alterations arise early in leukemogenesis, and alterations that drive signalling alterations are secondary events.

Similarly, analysis of the DNA methylation profiles of 27 cases of MPAL (11 with multiple subpopulations), 74 non-MPAL leukaemias and 17 normal progenitor samples showed distinct methylation profiles between leukaemia subtypes, but not between MPAL subclones (Extended Data Fig. 7b–e and Supplementary Table 37). Thus, cytosine methylation does not drive immunophenotypic heterogeneity in MPAL.

Phenotypic plasticity of MPAL

To further examine the basis of lineage plasticity in MPAL, we used xenograft models in which immunophenotypic subpopulations were purified and transplanted into immunocompromised NOD/SCID/IL2Rγ-null-3/GM/SF (NSG-SGM3) mice. Sorted subpopulations of cells from a patient with T/M MPAL (Fig. 3c and Extended Data Fig. 8a), the *ZNF384*r B/M JIH-5 cell line[15] (Extended Data Fig. 8b, c), and a patient with *KMT2A*r MPAL (Extended Data Fig. 8d), when transplanted into multiple independent NSG-SGM3 mice, propagated the immunophenotypic diversity of the primary samples. Moreover, we observed a phenotype shift in a sample from a patient with T/M MPAL during passaging of the bulk tumour sample, with engraftment of either a B/M or T/M leukemia phenotype (Extended Data Fig. 8e–h). These data demonstrate the multilineage potential of phenotypic subpopulations in MPAL, and phenotypic evolution even in the absence of therapeutic pressure.

Collectively, our genomic data and in vivo lineage plasticity data suggest that intra-sample lineage diversification in MPAL is driven by constellations of genomic alterations acquired in a haematopoietic stem or progenitor cell with multilineage potential. To test this idea, we purified progenitor cell and blast populations and normal mature lymphocytes from samples from a patient with *ZNF384*r B/M MPAL and two patients with *WT1*-altered T/M MPAL (Fig. 4a and Extended Data Fig. 9a, b). Alterations identified in the unfractionated samples (for example, *TCF3–ZNF384* and mutations in *MYCN*, *NTSD2* and *DNAH17* in the *ZNF384*r sample) were identified in the purified blast populations but not in non-leukaemic T or natural killer (NK) cells. Each alteration was also present in multiple haematopoietic progenitor populations with myeloid and lymphoid potential, and a subset of HSCs (Fig. 4b and Extended Data Fig. 9c). Analogous results were detected in two cases of T/M MPAL with *WT1* alterations (data not shown); these contrast with Ph-like B-ALL, in which founding lesions are detectable in a primitive progenitor with the capacity for myelo-lymphoid differentiation, but not in HSCs[16]. These data support the notion that mutations are acquired in a HSC that is primed for lineage aberrancy.

To gain further insight into the relative roles of founding genomic lesions, acquired genetic alterations and the role of therapy in dictating MPAL phenotype, we analysed sequential samples obtained at initial diagnosis and disease recurrence in nine patients. The immunophenotypes of five cases (three T/M MPAL, one B/M MPAL, and one MPAL NOS with T/B phenotype) were stable from diagnosis and relapse, but changed in four cases. Two were ALL (one B-ALL, one ETP-ALL) at diagnosis and relapsed as MPAL, and two were MPAL at diagnosis (one T/M, one B/M) and subsequently relapsed as AML and ALL, respectively (Extended Data Fig. 10). In the five cases with immunophenotypic stability, mutations in the predominant clone were lost (*PTPN11*, *CCND3*, *NOTCH1*, and *RPL22*) or emerged (*TP53*, *IKZF1*, *NF1*, *NCOR1*, and *SUZ12*). Despite this genomic evolution, the lineage ambiguity remained, further supporting the notion that MPAL leukaemia-initiating cells are primed for multi-lineage potential. In all four cases with phenotype shifts, the initial therapy correlated with the type of phenotype shift: patients who received ALL-directed therapy relapsed with myeloid leukaemia and one patient who received AML-directed therapy relapsed with lymphoid leukaemia. In two cases, immunophenotype at relapse was also correlated with a mutation characteristic of leukaemia subtype: *CEBPA* for AML and *CDKN2A* or *CDKN2B* for B-ALL. Together, these nine cases with serial samples support the theory that early genomic lesions prime progenitors for lineage aberrancy, which may remain stable or change over time, and

that phenotype is influenced by therapeutic pressure and/or genomic evolution.

Discussion

This study provides a comprehensive genomic analysis of paediatric MPAL, providing insights into the genomic relationships between immunophenotypically defined subtypes of acute leukaemia. We propose an update to the WHO classification of acute leukaemia that includes new subtypes of *ZNF384*-rearranged acute leukaemia (either B-ALL or MPAL), *WT1*-mutant T/M MPAL, and Ph-like B/M MPAL (Extended Data Fig. 1c).

The ALL-like genomic landscape of B/M MPAL and the similarity in genomic alterations between *ZNF384*r B/M MPAL and B-ALL supports the use of ALL-directed therapy for patients with B/M MPAL. Furthermore, the overexpression of FLT3 and responsiveness to FLT3 inhibition in *ZNF384* leukaemia[17] suggest that such targeted therapy should be considered in this form of leukaemia. Non-*ZNF384*r cases of B/M MPAL should be carefully evaluated for other kinase-activating alterations that may be amenable to kinase inhibition, as shown in Ph-like ALL[18].

Our data show that ETP-ALL[5] and T/M MPAL are genomically and epigenomically similar, and suggest that FLT3 and/or JAK inhibition should be evaluated further[4]. T/M MPAL exhibits infrequent alteration of core T-ALL transcription factor genes and few mutations in *CDKN2A*, *CDKN2B*, *NOTCH1* and *FBXW7*; frequent *FLT3*-activating mutations; and a GEP that overlaps with that of AML, consistent with the notion that the pathogenesis of T/M MPAL is distinct from that of T-ALL. However, contemporary paediatric ALL trials have demonstrated remarkable success in treating ETP-ALL, which is similar to T/M MPAL, so ALL-directed therapy may also be appropriate for T/M MPAL[19].

In contrast to the notion that subclonal genomic variation drives clonal evolution during disease progression in ALL[14], our analysis of phenotypically distinct subpopulations within individual patients with MPAL revealed that mutational variegation did not determine phenotypic diversification. Rather, the common genomic features of *ZNF384*r B-ALL and MPAL, limited mutational variegation between subclones, multi-lineage potential of subclones in xenograft models, lineage plasticity in serial patient samples, and identification of leukaemia-initiating alterations in early haematopoietic progenitors indicate that the ambiguous phenotype of MPAL is the result of the acquisition of alterations in immature haematopoietic progenitors (Fig. 4c, d). These data also support a model of haematopoiesis in which progenitors retaining multilineage potential undergo terminal differentiation into a single lineage only relatively late in haematopoiesis[20].

By demonstrating the genomic similarity of phenotypically distinct malignant populations, and by identifying the potential clinical importance of *ZNF384* fusions, these results emphasize the limitations of morphology and immunophenotype alone in diagnostic evaluation. As has been demonstrated in AML, ETP-ALL, myelodysplastic syndrome, and Ph-like ALL[5,11,18,21,22], accurate MPAL sub-classification requires careful genomic analysis to optimally guide diagnosis, risk-stratification and tailoring of therapy. Together, these findings have implications for disease classification and therapeutic decisions, while also clarifying the pathogenesis of this high-risk subtype of acute leukaemia.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0436-0.

Received: 1 September 2017; Accepted: 3 July 2018;
Published online 12 September 2018.

1. Gerr, H. et al. Acute leukaemias of ambiguous lineage in children: characterization, prognosis and therapy recommendations. *Br. J. Haematol.* **149**, 84–92 (2010).
2. Rubnitz, J. E. et al. Acute mixed lineage leukemia in children: the experience of St Jude Children's Research Hospital. *Blood* **113**, 5083–5089 (2009).
3. Matutes, E. et al. Mixed-phenotype acute leukemia: clinical and laboratory features and outcome in 100 patients defined according to the WHO 2008 classification. *Blood* **117**, 3163–3171 (2011).
4. Maude, S. L. et al. Efficacy of JAK/STAT pathway inhibition in murine xenograft models of early T-cell precursor (ETP) acute lymphoblastic leukemia. *Blood* **125**, 1759–1767 (2015).
5. Zhang, J. et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**, 157–163 (2012).
6. Swerdlow, S. H. et al. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues (Revised 4th Edition)*. (IARC, Lyon, 2017).
7. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
8. Yasuda, T. et al. Recurrent DUX4 fusions in B cell acute lymphoblastic leukemia of adolescents and young adults. *Nat. Genet.* **48**, 569–574 (2016).
9. Williams, R. T., Roussel, M. F. & Sherr, C. J. *Arf* gene loss enhances oncogenicity and limits imatinib response in mouse models of Bcr–Abl-induced acute lymphoblastic leukemia. *Proc. Natl Acad. Sci. USA* **103**, 6688–6693 (2006).
10. Coustan-Smith, E. et al. Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia. *Lancet Oncol.* **10**, 147–156 (2009).
11. Liu, Y. et al. The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat. Genet.* **49**, 1211–1218 (2017).
12. Bolouri, H. et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat. Med.* **24**, 103–112 (2018).
13. Mansour, M. R. et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).
14. Ma, X. et al. Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. *Nat. Commun.* **6**, 6604 (2015).
15. Ping, N. et al. Establishment and genetic characterization of a novel mixed-phenotype acute leukemia cell line with EP300–ZNF384 fusion. *J. Hematol. Oncol.* **8**, 100 (2015).
16. Iacobucci, I. et al. Truncating erythropoietin receptor rearrangements in acute lymphoblastic leukemia. *Cancer Cell* **29**, 186–200 (2016).
17. Griffith, M. et al. Comprehensive genomic analysis reveals FLT3 activation and a therapeutic strategy for a patient with relapsed adult B-lymphoblastic leukemia. *Exp. Hematol.* **44**, 603–613 (2016).
18. Roberts, K. G. et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *N. Engl. J. Med.* **371**, 1005–1015 (2014).
19. Conter, V. et al. Early T-cell precursor acute lymphoblastic leukaemia in children treated in AIEOP centres with AIEOP-BFM protocols: a retrospective analysis. *Lancet Haematol.* **3**, e80–e86 (2016).
20. Notta, F. et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016).
21. Lindsley, R. C. et al. Prognostic mutations in myelodysplastic syndrome after stem-cell transplantation. *N. Engl. J. Med.* **376**, 536–547 (2017).
22. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).

Acknowledgements We thank the Biorepository, the Genome Sequencing Facility of the Hartwell Center for Bioinformatics and Biotechnology, and the Flow Cytometry and Cell Sorting core facility and Cytogenetics core facility of St. Jude Children's Research Hospital (SJCRH). This work was supported in part by the American Lebanese Syrian Associated Charities of SJCRH, Cookies for Kids Cancer (to H.I.), St. Baldrick's Foundation Robert J. Arceci Innovation Award and Henry Schueler 41&9 Foundation (to C.G.M.), SJCRH Physician Scientist Training Program Fellowship (to T.B.A.), the National Cancer Institute grants P30 CA021765 (SJCRH Cancer Center Support Grant), Chair's grant and supplement to support the COG ALL TARGET project), U10 CA98413 (to the COG Statistical Center), U24 CA114766 (to COG; Specimen Banking), and Outstanding Investigator Award R35 CA197695 (to C.G.M.). The results published here are in part based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatments initiative of the NCI (http://ocg.cancer.gov/programs/target). This project has been funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract No. HHSN261200800001E (to C.G.M. and Michael Smith Genome Sciences Centre). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. We acknowledge Canada's Michael Smith Genome Sciences Centre, Vancouver, Canada for library construction and sequencing. A full list of funders of infrastructure and research supporting the services accessed is available at www.bcgsc.ca/about/funding_support.

Reviewer information *Nature* thanks R. Levine and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.B.A.: study design, flow analysis and sorting, data analysis, and manuscript writing. Z.G.: genomic data analysis. I.I.: genomic and mouse experiments, data analysis, data interpretation and manuscript preparation. K.D.: *ZNF384*r modelling. J.K.C.: central review of immunophenotype. B.X.: ChIP–seq and RNA-seq data analysis. D.P.-T. and H.Y.:

performed experiments. M.L.L. and S.P.H.: led and contributed to Children's Oncology Group ALL studies and the ALL TARGET project. M.B. and B.W.: reviewed flow cytometry. M.D., N.A.H., and A.C.: provided clinical data. J.H., E.O., B.B, G.B., S.E., V.d.H., C.M.Z., A.Y., D.R., D.T., N.K., T.L., B.D.M., D.C., H.H., A.M., A.S.M., O.H., K.E.N., J.R.D., and J.Z.: patient samples and clinical data. S.M.: data for comparison cohort. Y.-L.Y.: flow analysis. M.A.S., T.M.D., L.C.H., P.G., M.A.M., Y.M., A.J.M., R.A.M., S.J.M.J., and J.M.G.A.: genomic sequencing, analysis, and support. M.V.: performed FISH. L.J.J.: necropsy and histology on xenograft models. J.E.R. and C.-H.P.: patient samples and clinical data. D.S.G.: support for genomic analysis and manuscript editing. L.D. and Y.L.: genomic analysis. X.C., L.S., S.P. and D.P.: statistical analysis. S.N.: somatic and germline variant analysis. H.I.: acquisition of patient samples and clinical data. C.G.M.: designed and oversaw the study, analysed data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0436-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0436-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to H.I. or C.G.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Patients and samples. Diagnosis and remission samples were obtained from St. Jude Children's Research Hospital (SJCRH), the Children's Oncology Group, the European Organization for Research and Treatment of Cancer—Children's Leukaemia Group, the Belgian Society for Paediatric Hematology–Oncology, the Dutch Children's Oncology Group, the Italian Association of Paediatric Hematology and Oncology, the Japanese Association of Childhood Leukaemia Study, the Tokyo Children's Cancer Study Group, the I-BFM Study Group, the Queensland Children's Tumour Bank, The Children's Hospital at Westmead, Schneider Children's Medical Center, Yong Loo Lin School of Medicine in Singapore, and the United Kingdom Childhood Leukaemia Cell Bank. After central review of pathology and immunophenotyping of 159 cases, 115 patients diagnosed with ALAL were included in this analysis, including 80 with germline samples. We examined leukaemia samples from 115 patients with ALAL (Supplementary Tables 2–4) using whole-exome sequencing (WES) or whole-genome sequencing (WGS), transcriptome sequencing (RNA-seq), SNP microarray, and methylation array analysis. Samples collected on tumour banking protocols were used. Samples were not prospectively collected. The study was approved by the SJCRH Institutional Review Board.

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Tissue. Non-tumour DNA was extracted from remission bone marrow or peripheral blood samples, flow-sorted normal lymphocytes, or cultured fibroblasts using phenol–chloroform organic extraction. Tumour DNA was extracted using phenol–chloroform organic extraction. Tumour RNA was extracted using a TRIzol (Life Technologies).

WGS, WES and transcriptome sequencing. WGS for 44 cases and RNA-seq for 45 cases were performed by the British Columbia Cancer Agency's Michael Smith Genome Sciences Centre (BCGSC); WGS for 3 cases, WES for 92 cases and RNA-seq for 77 cases were performed at SJCRH. For WGS at BCGSC, methods for DNA preparation, sequencing, and quality control are available at https://ocg.cancer.gov/programs/target/target-methods. For WES at SJCRH, library construction used DNA tagmentation (fragmentation and adaptor attachment) performed using the reagent provided in the Illumina Nextera rapid exome kit, and was performed using the Caliper Biosciences (Perking Elmer) Sciclone G3. First-round PCR (10 cycles) was performed using Illumina Nextera kit reagents, and clean-up steps employ BC/Agencourt AMPure XP beads. Target capture used Illumina Nextera rapid capture exome kit and supplied hybridization and associated reagents. The pre-hybridization pool size was 12 samples, and second round PCR (10 cycles) performed with Nextera kit reagents. Library quality control was performed using a Victor fluorescence plate reader with Quant-it dsDNA reagents for pre-pool quantification, and Agilent Bio-analyzer 2200 for final library quantification. Paired-end sequencing was performed using Illumina HiSeq 2500 with read length 100 bp.

Methods for RNA preparation, sequencing, and quality control at BCGSC are available at https://ocg.cancer.gov/programs/target/target-methods. At SJCRH, total RNA quality and quantity were assessed on Agilent RNA6000 chips (Agilent Technologies) and Qubit (Life Technologies). RNA-seq libraries were prepared from 500 ng of total RNA for each sample following Illumina RNA-seq protocols, including DNase treatment and phenol purification, cDNA conversion, fragmentation by Covaris Ultrasonicator, end repair, deoxyadenosine tailing, adaptor ligation and PCR amplification (ten cycles). Libraries with a 10 pM concentration were clustered on an Illumina cBot, and each flow cell was loaded onto a HiSeq instrument for sequencing using the Illumina 2×100 bp sequencing kit. RNA-seq was not performed on flow-sorted subpopulations due to the deleterious effects on RNA integrity of cellular fixation/permeabilization performed to enable staining for intracellular markers.

Sequencing read alignment. Paired-end WGS and WES data were aligned to the human reference genome GRCh37 by BWA[23] (version 0.7.12). Samtools[24] (version 1.3.1) was used to generate chromosomal coordinate-sorted and indexed BAM files, and then the Picard (http://broadinstitute.github.io/picard/, version 1.129) MarkDuplicates module was used for marking PCR duplication. Afterwards, the reads were realigned around potential indel regions by GATK[25] (version 3.5) IndelRealigner module following the recommended pipeline. Sequencing depth and coverage was evaluated based on coding regions defined by refSeq genes from UCSC, with the length around 34 Mb.

SNV/indel calling and filter workflow. The GATK UnifiedGenotyper module was used to identify SNVs and indels from leukaemia and germline samples, which were filtered by a homemade pipeline, excluding: 1) reported common SNPs/indels from UCSC dbSNP v142; 2) germline mutations detected from matched germline control samples. All the non-silent SNVs/indels yield from the filtering pipeline were manually reviewed and only the highly reliable somatic ones were reported. Meanwhile, adjacent nucleotide changes on the same allele were merged into a single mutation. For patients with flow-sorted subpopulations of sequenced leukaemia cells, the mutation calling for each population was performed de novo. Mutations detected from some/one of the samples were checked across the other samples from the same patient. In these cases, we applied a threshold of at least 3 mutant allele reads and VAF of at least 1% to report a mutation. For cases without germline samples, a germline sample was picked with highest sequence depth as a pseudo-germline sample to run through the filtering pipeline. In cases in which flow-sorted subpopulations were sequenced, WES or WGS of the unfractionated samples were not performed.

Structure variant detection. Structural variants in the tumours were identified by CREST[26] using the tumour vs germline mode, with pseudo-germline data applied for tumours without germline samples. Candidate variants were manually reviewed and the mapping uniqueness was re-evaluated by running BLAT[27] mapping and the confident calls were considered as the final structural variant set.

RNA-seq data analysis for patient samples. Paired-end reads were mapped to the GRCh37 human genome reference by STAR[28] (version 2.5.1b) through the recommended two pass mapping pipeline with default parameters and the Picard MarkDuplicates module was used to mark the duplication rate. Gene annotation files were downloaded from Ensembl (http://www.ensembl.org/) and used for STAR mapping and subsequent gene expression level evaluation. CICERO[18] and FusionCatcher[29] were used to detect fusions from mapped BAM files and raw FASTQ files, respectively. The reported fusion contigs were remapped by BLAT to check the reliability of mapping quality, the breakpoints were manually reviewed from the aligned reads and the highly confident fusions were reported. To evaluate GEP, reads count for annotated genes was called by HTSeq[30] (version 0.6.0) and processed by DESeq2 R package[31] to normalize gene expression into regularized $\log_2$ values (rlog). Six cases without DNA-sequence data were screened for SNVs/indels by following the GATK best practices for variant calling on RNAseq (https://gatkforums.broadinstitute.org/gatk/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail). The filtering process is the same as for germline variant analysis described below.

Gene set enrichment and pathway analysis. Read counts from RNA-seq data were imported to DESeq2[32] R package for differential gene expression analysis. To perform gene set enrichment analysis (GSEA)[33], all the genes were ranked according to the fold-change and significance from differential analysis. GSEA was performed using mSigDB C2 genes and curated gene sets from in-house analyses.

Cell line transcriptome analysis. Total RNA was isolated from green fluorescent protein (GFP)-positive, sorted cells using the RNeasy Mini Kit (Qiagen). RNA quality was checked using 2100 Bioanalyzer RNA 6000 Nanoassay (Agilent) or LabChip RNA Pico Sensitivity assay (PerkinElmer) before library generation. Libraries were prepared from total RNA with the TruSeq Stranded Total RNA Library Prep Kit (Illumina). Libraries were quantified using the Quant-iT PicoGreen dsDNA assay (Life Technologies) Kapa Library Quantification kit (Kapa Biosystems) or low pass sequencing on a MiSeq Nano v2 run (Illumina). One hundred cycle paired end sequencing was performed on an Illumina HiSeq 2500, HiSeq 4000, or NovaSeq 6000. RNA isolation, library preparation, and sequencing were performed on three biological replicates. RNA-seq data were mapped as described previously[18] and HTSeq[30] (version 0.6.1p1) were used to get gene-level count and estimated FPKM based on GENCODE (vM9)[34]. Voom[35] was used for gene differential expression analysis after trimmed mean normalization.

CNA and loss of heterozygosity (LOH). DNA from leukaemia and matched germline samples was prepared for hybridization to Illumina Infinium Omni2.5 Exome-8 SNP arrays according to the manufacturer's protocol. The raw intensity data (*.idat files) were analysed by the Genotyping Module of Illumina Genome Studio software version 2.0.3. Normalized log R ratio (LRR) and B allele frequency (BAF) for all the available probes in each sample were extracted. For *ZNF384*r B-ALL cases, data acquired from Affymetrix Genome-Wide Human SNP Array 6.0 were also converted to LRR and BAF values following the pipeline described by PennCNV[36] (http://penncnv.openbioinformatics.org/en/latest/user-guide/affy/). With the input of LRR and BAF, somatic genomic alterations in paired or unpaired samples were called by OncoSNP version 2.1[37]. To verify the reliability of CNAs and LOHs, all the reported alterations were plotted based on LRR and BAF in ShinyCNV (https://github.com/gzhmat/ShinyCNV) and visually checked[38]. Only somatic alterations meeting the criteria proposed by OncoSNP and PennCNV were kept for further analysis.

DNA methylation assay and data analysis. We examined DNA methylation profiles in 27 MPAL cases (11 with 2–4 subpopulations), 15 AML, 29 B-ALL, 30 T-ALL, and 17 normal lymphocyte samples from 4 healthy donors. Raw data from the Infinium MethylationEPIC BeadChip Kit (Illumina Inc.) were analysed using the ChAMP[39] R package. In general, the raw *.idat files were imported through 'minfi' method[40] and then the following filters were applied to exclude the probes: (1) with detection *P* value above 0.01 in one or more samples; (2) with beadcount <3 in at least 5% of samples; (3) as non-CpG probes; (4) identified as SNPs[41]; (5) aligned to multiple locations[42]; and (6) on the X or Y chromosome. After filtering, 'BMIQ' normalization from ChAMP package was used as the author suggested to calculate methylation beta values. Batch effect was observed by the singular

value decomposition method[43] and adjusted by ComBat normalization method[44]. The 5,000 probes with the highest median absolute deviations (MAD) were used to perform clustering with a two-dimensional *t*-distributed stochastic neighbour embedding (*t*-SNE) plot and heat map[45].

Fusion validation. Fluorescence in situ hybridization (FISH) was performed to confirm fusions in 22 cases (Supplementary Table 19) using the listed probes, in Carnoy's fixative as previously described[46]. BAC clones (Supplementary Table 20) were labelled with rhodamine or fluorescein isothiocyanate. At least 100 interphase nuclei were scored per case.

Flow cytometry analysis and flow cytometry-assisted cell sorting. Flow cytometry analysis and sorting were performed on an 18-colour Aria cell sorter (BD Biosciences). When available, cryopreserved samples were analysed by flow cytometry using CD45-APC-H7 (BD 560178), cytoplasmic CD3-PE (BD 347347), CD34-PerCP Cy5.5 (BD 347203), CD19-APC (BD 340437), cytoplasmic MPO-FITC (Dako F071401-1), and CD33-PE-Cy7 (BD 333946). Depending on the phenotypes reported from the outside institutions, samples were additionally analysed using cytoplasmic CD79a-APC (BD 551134), CD22-BV421 (BD 563940), CD64-PerCP-Cy5.5 (BD 561194), CD14-PE-Cy7 (BD 560919), cytoplasmic lysosome-FITC (Life Technologies GIC207), and CD11c-APC (BD 560895). For 50 cases, leukaemic cells in the CD45 and side scatter-defined blast gate were sorted into subpopulations based upon cytoplasmic MPO and either CD19 or cytoplasmic CD3. When feasible, normal lymphocytes were sorted using side scatter, CD45 lymphocyte gate and secondarily using CD19 and cytoplasmic CD3 to collect normal B cells in T/M MPAL cases and normal T cells in B/M, *KMT2A*r, AUL, or NOS cases.

Fibroblast cultures. Bone marrow cells were cultured in change medium (Irvine scientific, T105), which was changed every 5 days. Cells were collected for DNA extraction when the fibroblasts became at least 70% confluent.

Comparison cohorts. Comparison cohorts of AML, ETP-ALL, non-ETP T-ALL, and B-ALL were examined. A cohort of 197 paediatric patients with AML from the COG with WGS performed though the NCI TARGET initiative was used as comparison (Supplementary Table 34) and publicly available data can be found at https://ocg.cancer.gov/programs/target[12]. Nineteen ETP-ALL and 245 non-ETP T-ALL cases from the COG were sequenced through the NCI TARGET project using WES and total stranded RNA-seq for SNV, indel, CNA and SV calls, fusion detection and GEP comparison[11] (Supplementary Table 35). A cohort of AML, B-ALL[16,18,47–49] ($n=161$), T-ALL[11] ($n=50$), ETP-ALL[11] ($n=19$) and 12 normal lymphocyte samples was used for GEP comparison (Supplementary Table 23). The AML samples were sequenced at SJCRH and had stranded total RNA-seq for GEP comparisons. This cohort consists of five cases with core binding factor translocations (three with *RUNX1–RUNX1T1*, two with *CBFB–MYH11*), five cases with normal karyotype, and five cases with *KMT2A*r).

B-ALL subtyping based on GEP. RNA-seq data analysis for patient samples is described above. As many B-ALL subtypes defined by single chromosomal aneuploidy or rearrangement may be clustered based on their GEP[48,50–54], a subtype predication model was trained by prediction analysis of microarrays (PAM)[55] using a cohort of 322 B-ALL samples from our previous studies[18,48,49], which consists of eight canonical B-ALL subtypes: *DUX4* rearrangement ($n=40$), *ETV6–RUNX1* ($n=42$), high hyperdiploidy ($n=45$), *MEF2D* rearrangement ($n=29$), *KMT2A*r ($n=44$), *TCF3–PBX1* ($n=40$), *BCR–ABL1* ($n=42$) and *ZNF384* rearrangement ($n=40$). The PAM model was trained on 200 different thresholds with tenfold cross-validation. On the basis of the trained model and cross-validation result, 100 thresholds (control the selected feature genes from 5,000 to 50) were tested on the training data set to determine the optimal threshold range for each subtype. Then the trained model was applied to the MPAL samples to determine their similarity to each B-ALL subtype for 100 rounds, using evenly distributed thresholds across the optimal threshold range for each B-ALL subtype, and the average score was taken as the consensus likelihood score for that subtype (Supplementary Table 33).

Germline variant analysis. Germline variants were called by GATK[56] UnifiedGenotyper from the BAM files of all the germline samples, and then the following filters were applied to identify potential pathogenic germline variants: (1) exclusion of variants with fewer than five mutant reads support or a VAF below 20%; (2) exclusion of variants in common SNP database (VAF greater than 0.1% in population according to dbSNP 142); (3) exclusion of SNPs with at least ten occurrences observed in dbSNP 142 but not reported as somatic mutations in the COSMIC V80 database; (4) exclusion of variants in genes with fewer than three somatic mutations in MPAL cohort; (5) annotation of variants using the Variant Effect Predictor (VEP; https://useast.ensembl.org/Tools/VEP) and then exclusion of variants predicted as benign by any of the predictors (SIFT, PolyPhen, Condel). The remaining mutations were manually reviewed and obvious mapping artefacts were excluded. Mutations were then assessed according to ACMG recommendations[7].

Lentiviral transduction of cells. cDNAs encoding ZNF384 (XM_017018949), TAF15 (NM_139215)–ZNF384 (exon 6–exon 3), and TCF3 (NM_003200)–ZNF384 (exon 13–exon 5) were amplified from human leukaemic cell RNA and cloned with a C-terminal HA epitope tag (added using the QuikChange II XL Site-Directed Mutagenesis Kit, Agilent) into the CL20c-MSCV-IRES-GFP vector. Vectors were packaged into lentiviral particles by transient transfection of HEK293T cells with a triple plasmid (pHDMG, pCAG HIV, pCAG RTR) system. Lentiviral supernatants were used to infect interleukin-7 (IL-7)-dependent *Arf*$^{-/-}$ pre-B cells on RetroNectin (Takara Bio) for 48 h before sorting for GFP$^+$ cells (BD FACSAria, BD Biosciences).

Chromatin immunoprecipitation and sequencing. ChIP assays were carried out as described previously[49]. In brief, 2×10^7 GFP-positive cells were incubated for 10 min in 1% formaldehyde in phosphate-buffered saline (PBS) at room temperature, quenched by the addition of 1/10 volume of 2 M glycine. Cells were then washed three times with cold PBS containing proteinase inhibitors and lysed on ice for 10 min in lysis buffer (50 mM HEPES, pH 7.9, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100). Chromatin was washed twice in washing buffer (10 mM Tris-HCl, pH 8, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA) and then twice in shearing buffer (0.1% SDS, 10 mM Tris-HCl, pH 8, 1 mM EDTA) before resuspension in 1 ml shearing buffer. Chromatin was sonicated in 1-ml AFA millitubes using a Covaris E210 instrument for 15 min at 5% duty cycle, intensity 4, 200 cycles per burst at 4 °C. Sheared chromatin was spun down for 10 min at 13,200*g* at 4 °C, and the supernatant was mixed with an equal amount of ChIP dilution buffer (0.1% SDS, 30 mM Tris-HCl, pH 8, 1 mM EDTA, 300 mM NaCl, 2% Triton X-100) before ChIP experiments. Immunoprecipitation was performed with an antibody to HA (ab9110, Abcam) and a normal rabbit IgG control (Santa Cruz Biotechnology) using 2 μg antibody per ChIP. This experiment was performed with three biological replicates.

To prepare ChIP–seq libraries, 10 ng of ChIP DNA was end repaired and adaptor ligation was performed using the Next ChIP–Seq Library Prep Reagent Set for Illumina (New England BioLabs). Libraries were purified after 14 rounds of PCR amplification with Q5 DNA Hot-Start polymerase (New England BioLabs). Each ChIP–seq library underwent 50-cycle single-end sequencing using TruSeq SBS kit v3 on an Illumina HiSeq 2000.

Alignment and quality control were performed as described[57]. Fifty base pair single-end reads were mapped to mouse genome mm9 (MGSCv37) with BWA[23] (version 0.7.12-r1039), duplicated reads were marked with Picard and only unique mapped reads extracted by Samtools[24] (version 1.2) were kept for analysis. We extended each read to estimated fragment size by SPP[58] (version 1.1) and generated bigwig files, scaling the track by normalizing to 15 million unique mapped reads.

For differential binding analysis, peaks were called with MACS2[59] (version 2.0.10.20131216, parameter '–nomodel–extsize FRAGMENT SIZE' and fragment size was estimated as described above by SPP[58] (version 1.1) twice for each sample. High confidence peaks used a cutoff of FDR-corrected *P* value of 0.05 and low confidence peaks used a cutoff of FDR-corrected $P<0.5$. Peaks from replicates were merged only if called as high confidence peaks in one sample and called as at least low confidence peaks in other replicates. Finally, peaks from wild-type *ZNF384* and *ZNF384* fusions were merged as a reference peak set. For each sample, we first extend read to the estimated fragment size, then we counted the extended reads number overlapping the reference peaks by BEDTools (version 2.24.0)[60]. Following PCA analysis, which showed a clear separation of wild-type and fusion ChIP–seq data, Voom[35] was used to examine differences in strength of binding between wild type and fusion after trimmed mean normalization. Common differential binding sites (*q* value less than 0.05 and fold change greater than 1) between TAF15–ZNF384 versus wild-type proteins and TCF3–ZNF384 versus wild-type proteins were used for visualization. Real-time PCR (ΔC_t method) was employed to validate ChIP–seq results. Differential binding sites were annotated to genes if their promoter (transcription start site ± 2 kb) overlapped the binding sites. GSEA[33] was used to compare ChIP–seq peak lists to the GEP of cell lines expressing ZNF384 fusions.

Statistical analysis. The correlation between sex, disease subtype (WHO 2016 criteria, our proposed update to classficiation of ALAL, or fusion presence/absence) and single gene mutation or pathway mutations was assessed using the two-sided Fisher exact test. The correlation between subtypes and age categories was assessed using the two-sided Fisher exact test. The correlation between age as a continuous variable and single gene mutation or pathway mutations was assessed using the non-parametric Wilcoxon rank-sum test. The Kaplan–Meier method was used to estimate the survival function and overall survival distributions were compared with log-rank tests. GraphPad Prism (version 7.04) and SAS (version 9.4) were used for statistical analysis.

Fluorescence-activated cell sorting (FACS) of human stem/progenitor and mature cell populations. For sorting of HSC and progenitor cells, mononuclear cells from diagnosis bone marrow samples from patient SJMPAL040028 were stained with the following human-specific antibodies (all from BD Biosciences unless stated otherwise, catalogue number in parentheses): anti-CD45RA-FITC (555488), anti-CD90-PE (Biolegend, 328109), anti-CD135-BV711 (563908),

anti-CD38-PE-Cy7 (335790), anti-CD10-BV421 (562902), anti-CD7-V450 (642916), anti-CD45-AlexaFluor 700 (Thermo Fisher Scientific MHCD4529), anti-CD34-APC-Cy7 (Biolegend, custom-made, CD34 clone 581), anti-CD33-APC (340680) and anti-CD19-BV605 (562653). For sorting of mature cells and leukaemic blasts, mononuclear cells from bone marrow of patient SJMPAL040028 were stained with the following antibodies: anti-CD45-AlexaFluor 700 (Thermo Fisher Scientific MHCD4529), anti-CD19-BV605 (562653), anti-CD10-BV421 (562902), anti-CD33-PE-Cy7 (333946), CD3-PE (347347) and anti-CD56-AlexaFluor 647 (557711). For all samples, cells (from 5 to 1,000) per fraction were sorted on a BD FACS Aria in a 96-well plate. As previously published[16] and as described[61], progenitor populations were all gated on $CD45^{+}CD33^{-}CD19^{-}$ and sorted into HSCs ($CD38^{-}CD34^{+}CD90^{+}CD45RA^{-}$); multipotent progenitor fraction (MPP; $CD38^{-}CD34^{+}CD90^{-}CD45RA^{-}$); multilymphoid progenitor fraction (MLP; $CD38^{-}CD34^{+}CD45RA^{+}$); megakaryocyte erythroid progenitors (MEP)/common myeloid progenitors (CMP; $CD38^{+}CD34^{+}CD7^{-}CD10^{-}CD45RA^{-}$); and granulocyte monocyte progenitor (GMP; $CD38^{+}CD34^{+}CD7^{-}CD10^{-}CD45RA^{+}$) subsets. Leukaemia blasts were gated on $CD45^{dim}$ expression and sorted into the following fractions: $CD45^{dim}CD33^{+}CD19^{+}CD10^{-}$; $CD45^{dim}CD33^{+}CD19^{moderate}CD10^{-}$; $CD45^{dim}CD33^{+}CD19^{-}CD10^{-}$; and $CD45^{dim}CD33^{-}CD19^{-}$. Normal mature populations were gated on $CD45^{high}$ expression and sorted into T cells ($CD45^{high}CD3^{+}$) and NK cells ($CD45^{high}CD56^{+}$). The following numbers of cells were sorted in a single well of a 96-well plate (each number in the parenthesis is a replicate): HSC (6 and 21); MPP (387); MLP (12); MEP/CMP (500); GMP (21 and 18); $CD45^{dim}CD33^{+}CD19^{+}CD10^{-}$ (1,000; 5 replicates); $CD45^{dim}CD33^{+}CD19^{moderate}CD10^{-}$ (1,000; 6 replicates); $CD45^{dim}CD33^{+}CD19^{-}CD10^{-}$ (1,000; 6 replicates); $CD45^{dim}CD33^{-}CD19^{-}$ (82 and 36); T cells (1,000; 5 replicates) and NK cells (100 and 40). DNA from all sorted populations was amplified by whole-genome amplification (WGA) by REPLI-g Single Cell Kit (150345, Qiagen) according to the manufacturer's protocol.

Genomic analysis of sorted subpopulations. Upon completion of WGA, DNA was subjected to PCR amplification using primers specific for the *TCF3–ZNF384* fusion or for additional genetic alterations, including SNVs/indels in *NDST2*, *DNAH17* and *MYCN*, identified from analysis of WES data. Primers were designed to flank the fusion breakpoint or the identified variants using Primer3 (*TCF3–ZNF384* forward: 5′-GAGGAGGACCAGGAGAGATGG-3′ and *TCF3–ZNF384* reverse: 5′-ATCAGGCAAGGCTTCCTAAAAG-3′; *NDST2* forward: 5′-ATAGGTACACTCCCTGCCTTTCC-3′ and *NDST2* reverse: 5′-ACCCCAAACCTTGACCCTTTT-3′; *DNAH17* forward: 5′-CTCCTC TTTGGGAACCCTCTG-3′ and *DNAH17* reverse: 5′-GAAAAGGCTTG TGCTGACATCTT-3′; *MYCN* forward: 5′-GTGTCTGTCGGTTGCAGTGTT-3′ and *MYCN* reverse: 5′-AGCTCGTT CTCAAGCAGCATCT-3′). PCR was performed using KAPA2G Fast HotStart Ready Mix (07961260001, Kapa Biosystems) according to the manufacturer's instructions with 10 μM each primer and 2 μl diluted (1:100) WGA DNA. All amplicons were quality checked on a 1.5% agarose gel and purified using Wizard SV Gel and PCR Clean-Up System (A9282, Preprotech). Sequences were verified by Sanger sequencing. The sequenced amplicon was aligned to a reference fusion sequence generated from National Center for Biotechnology Information and to the contigs obtained from RNA-seq in case of *TCF3–ZNF384* fusion. The results were analysed using CLC Main Workbench (Qiagen).

Xenografts. Mice were housed in an American Association of Laboratory Animal Care (AALAC)-accredited facility and were treated according to Institutional Animal Care and Use Committee (IACUC) protocols approved by SJCRH in accordance with NIH guidelines.

JIH-5-derived xenograft. The JIH-5 cell line[15] was obtained from Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ). Cells were thawed and cultured according to DSMZ's instructions (https://www.dsmz.de/catalogues/details/culture/ACC-788.html). Immunophenotypic and genomic analyses (RNA-seq) were performed prior transplant assays. Short tandem repeat (STR) DNA analysis was performed for cell line authentication (Supplementary Table 38), showing concordance with DSMZ STR analysis. STR analysis was performed using the The PowerPlex 16 HS System (Promega) which allows co-amplification and three-colour (blue or fluorescein-labelled, black or TMR-labelled, and green or JOE-labelled) detection of sixteen loci (fifteen STR loci and Amelogenin), including Penta E, D18S51, D21S11, TH01, D3S1358, FGA, TPOX, D8S1179, vWA, Amelogenin, Penta D, CSF1PO, D16S539, D7S820, D13S317 and D5S818. All sixteen loci were amplified simultaneously in a single tube and analysed in a single injection. Cells were transduced with a lentiviral vector (vCL20SF2-Luc2a-YFP) expressing luciferase and yellow fluorescent protein (YFP) and FACS sorted for YFP. YFP-positive (YFP^{+}) cells were stained with anti-CD19-APC (BD, 340437), anti-CD34 PerCP-Cy5.5 (BD, 347203) and anti-CD33 PE-Cy7 (BD, 333946) and sorted in the following subpopulations: $YFP^{+}CD34^{+}$; $YFP^{+}CD34^{-}CD19^{+}CD33^{+}$; and $YFP^{+}CD34^{-}CD19^{+}CD33^{-}$. FACS-sorted leukaemia subpopulations or YFP^{+} bulk cells were intravenously injected in 8- to 10-week-old female NSG-SGM3 mice[62]. The number of cells that was transplanted and the total number of mice transplanted per subpopulation depended on the number of viable cells that were available, but ranged from 0.2 to 0.6 million cells and 1 to 5 mice, respectively. Engraftment was monitored by weekly measurement of bioluminescence (region of interest (ROI)) at Xenogen IVIS-200 (PerkinElmer). ROI measurements and total fluxes (photons/second, p/s) were recorded and analysed by the Living Imaging v.4.4 software (Caliper Life Sciences). When total fluxes were at least 1×10^{8} in all animals, mice were euthanized, and blood, bone marrow, and spleen samples were analysed to determine the leukaemia phenotype, using morphology, flow cytometry, and histopathologic analysis.

MPAL patient-derived xenografts (PDX). MPAL PDX were established from three patients (SJMAPL011911, SJMAPL014124 and SJMAPL040036). Frozen mononuclear cells from bone marrow at diagnosis were thawed and used as bulk (SJMAPL040036) or flow-sorted in transplantation assays. Cells from SJMAPL011911 were stained with the following human-specific antibodies: anti-CD45-APC-H7 (BD, 641399), CD34-PerCP-Cy5.5 (BD, 347203), anti-CD33-PE-Cy7 (BD, 3339460) and anti-CD7-PE-Cy7 (BD, 544019). Blast cells were gated on $CD45^{dim}$ expression and sorted into $CD45^{dim}CD7^{+}CD33^{-}$, $CD45^{dim}CD7^{-}CD33^{+}$ and $CD45^{dim}CD7^{-}CD33^{-}$. Cells from SJMAPL014124 were stained with the following human-specific antibodies: anti-CD45-APC-H7 (BD, 641399), anti-CD33 PE-Cy7 (BD, 333946) and anti-CD19-APC (BD, 340437) and sorted into $CD45^{dim}CD19^{+}CD33^{+}$, $CD45^{dim}CD19^{+}CD33^{-}$ and $CD45^{dim}CD19^{-}CD33^{+}$. Bulk or sorted cells were intravenously injected into 8- to 10-week-old female NSG-SGM3[62] mice that were sublethally irradiated (250 RAD) 6–24 h before transplantation. The number of cells that was transplanted and the total number of mice transplanted per sample depended on the number of viable cells that were available, but ranged from 0.2 to 0.6 million cells and 1 to 5 mice, respectively. Human leukaemia engraftment was monitored in peripheral blood by performing serial retro-orbital bleeds one month after injections and monthly thereafter. Peripheral blood samples were analysed by flow cytometry for human $CD45^{+}$ cells and when $CD45^{+}$ cells were >5%, mice were euthanized, and blood, bone marrow, and spleen samples were analysed to determine the leukaemia phenotype, using morphology, flow cytometry, and histopathologic analysis. Immunohistochemistry (IHC) was performed on formalin-fixed paraffin-embedded tissues sectioned at 4 μm. Assays for CD19 (AbDserotec, MCA2454T; 1:100), CD34 (Ventana, 790-2927; ready to use), CD45 (Ventana, 760-2505; ready to use) and myeloperoxidase (MPO, DAKO A398; 1:500) were performed on the Ventana Benchmark. The assay for CD33 (Leica Biosystems, NCL-L-CD33; 1:200) was performed on the Dako Omnis.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Sequencing, SNP, and methylation data are available at the NCI Genomics Data Commons (GDC, gdc.cancer.giv) and analysed data may be accessed at the TARGET website at https://ocg.cancer.gov/programs/target/data-matrix or https://gdc.cancer.gov/about-data/publications/TARGET-ALAL-2018. Mouse RNA-seq and ChIP–seq data have been deposited in the GEO database under accession ID GSE112561. For T-ALL and ETP-ALL, RNA sequencing for comparison comprised previously published data[11]. B-ALL RNA-sequencing data for comparison comprised previously published data and recently sequenced samples that will be made available through St Jude's Children's Research Hospital[11,18,48,49,63]. T-ALL, ETP-ALL, and AML data for mutation comparison comprised previously published data[11,12]. The genomic landscape reported in this study can be explored at the St. Jude PeCan Data Portal, http://pecan.stjude.org/proteinpaint/study/pediatric-mpal.

23. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
24. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
25. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
26. Wang, J. et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8**, 652–654 (2011).
27. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
28. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
29. Edgren, H. et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.* **12**, R6 (2011).
30. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
31. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
32. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
33. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).

34. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
35. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
36. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
37. Yau, C. et al. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol.* **11**, R92 (2010).
38. Gu, Z. & Mullighan, C. G. ShinyCNV: a Shiny/R application to view and annotate DNA copy number variations. *Bioinformatics* https://doi.org/10.1093/bioinformatics/bty546 (2018).
39. Morris, T. J. et al. ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics* **30**, 428–430 (2014).
40. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
41. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).
42. Nordlund, J. et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol.* **14**, r105 (2013).
43. Teschendorff, A. E. et al. An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One* **4**, e8274 (2009).
44. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
45. van der Maaten, L. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
46. Mullighan, C. G. et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).
47. Andersson, A. K. et al. The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nat. Genet.* **47**, 330–337 (2015).
48. Gu, Z. et al. Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. *Nat. Commun.* **7**, 13331 (2016).
49. Zhang, J. et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. *Nat. Genet.* **48**, 1481–1489 (2016).
50. Den Boer, M. L. et al. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *Lancet Oncol.* **10**, 125–134 (2009).
51. Mullighan, C. G. et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *N. Engl. J. Med.* **360**, 470–480 (2009).
52. Lilljebjörn, H. et al. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat. Commun.* **7**, 11790 (2016).
53. Harvey, R. C. et al. Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome. *Blood* **116**, 4874–4884 (2010).
54. Yeoh, E. J. et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1**, 133–143 (2002).
55. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA* **99**, 6567–6572 (2002).
56. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
57. Aldiri, I. et al. The dynamic epigenetic landscape of the retina during development, reprogramming, and tumorigenesis. *Neuron* **94**, 550–568 (2017).
58. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP–seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).
59. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
60. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
61. Shlush, L. I. et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547**, 104–108 (2017).
62. Wunderlich, M. et al. AML xenograft efficiency is significantly improved in NOD/SCID-IL2RG mice constitutively expressing human SCF, GM-CSF and IL-3. *Leukemia* **24**, 1785–1788 (2010).
63. Roberts, K. G. et al. High frequency and poor outcome of Philadelphia chromosome-like acute lymphoblastic leukemia in adults. *J. Clin. Oncol.* **35**, 394–401 (2017).
64. Arber, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
65. Jaatinen, T. et al. Global gene expression profile of human cord blood-derived $CD133^+$ cells. *Stem Cells* **24**, 631–641 (2006).
66. Novershtern, N. et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).
67. Flotho, C. et al. Genes contributing to minimal residual disease in childhood acute lymphoblastic leukemia: prognostic significance of CASP8AP2. *Blood* **108**, 1050–1057 (2006).
68. Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).

a **Acute Leukemia of Ambiguous Lineage Subtypes**

Mixed Phenotype Acute Leukemia - Leukemia blasts express specific antigens from multiple leukocyte lineages
- **B/myeloid, NOS** - Leukemia blasts express both B-lymphoid and myeloid antigens
- **T/myeloid, NOS** - Leukemia blasts express both T-lymphoid and myeloid antigens
- ***KMT2A* rearranged** - Leukemia blasts express antigens from multiple lineages AND the presence of a *KMT2A* rearrangement
- ***BCR-ABL* positive** - Leukemia blasts express antigens from multiple lineages AND the presence of a *BCR-ABL* fusion
- **Not otherwise specified** - Leukemia blasts express both B and T-lymphoid antigens OR T-lymphoid, B-lymphoid, and myeloid antigens without a recurrent genetic abnormality

Acute Undifferentiated Leukemia - Leukemia blasts do not express any lineage defining antigens

b **Criteria for Lineage Assignment in Mixed Phenotype Acute Leukemia**

Myeloid Lineage
Myeloperoxidase
or
Moncytic differentiation (at least 2 of the following: NSE, CD11c, CD14, CD64, lysozyme)

T-lymphoid lineage
Cytoplasmic CD3 (flow cytometry with antibodies to CD3 epsilon chain; immunohistochemistry using polyclonal anti-CD3 antibody may detect CD3 zeta chain, which is not T-cell specific)
or
Surface CD3 (rare in mixed phenotype acute leukemia)

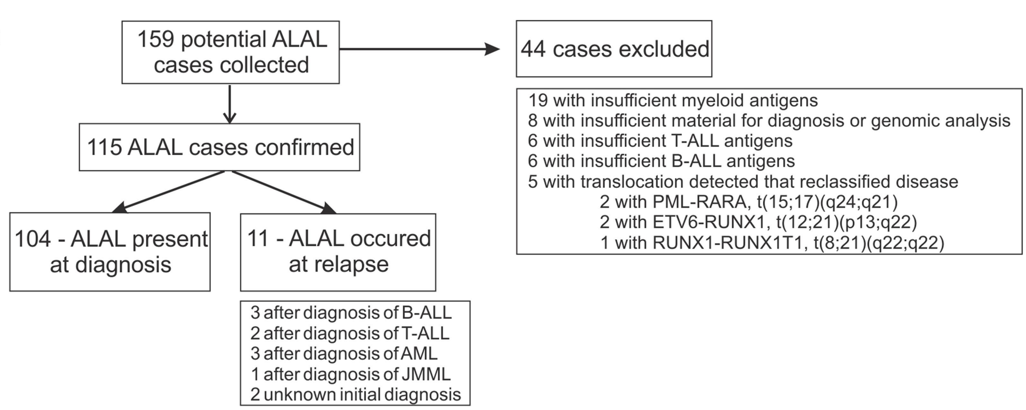
B-lymphoid lineage (multiple antigens required)
Strong CD19 with at least 1 of the following strongly expressed: CD79a, cytoplasmic CD22, CD10
or
Weak CD19 with at least 2 of the following strongly expressed: CD79a, cytoplasmic CD22, CD10

c **Proposed Update to Acute Leukemia of Ambiguous Lineage Subtypes**

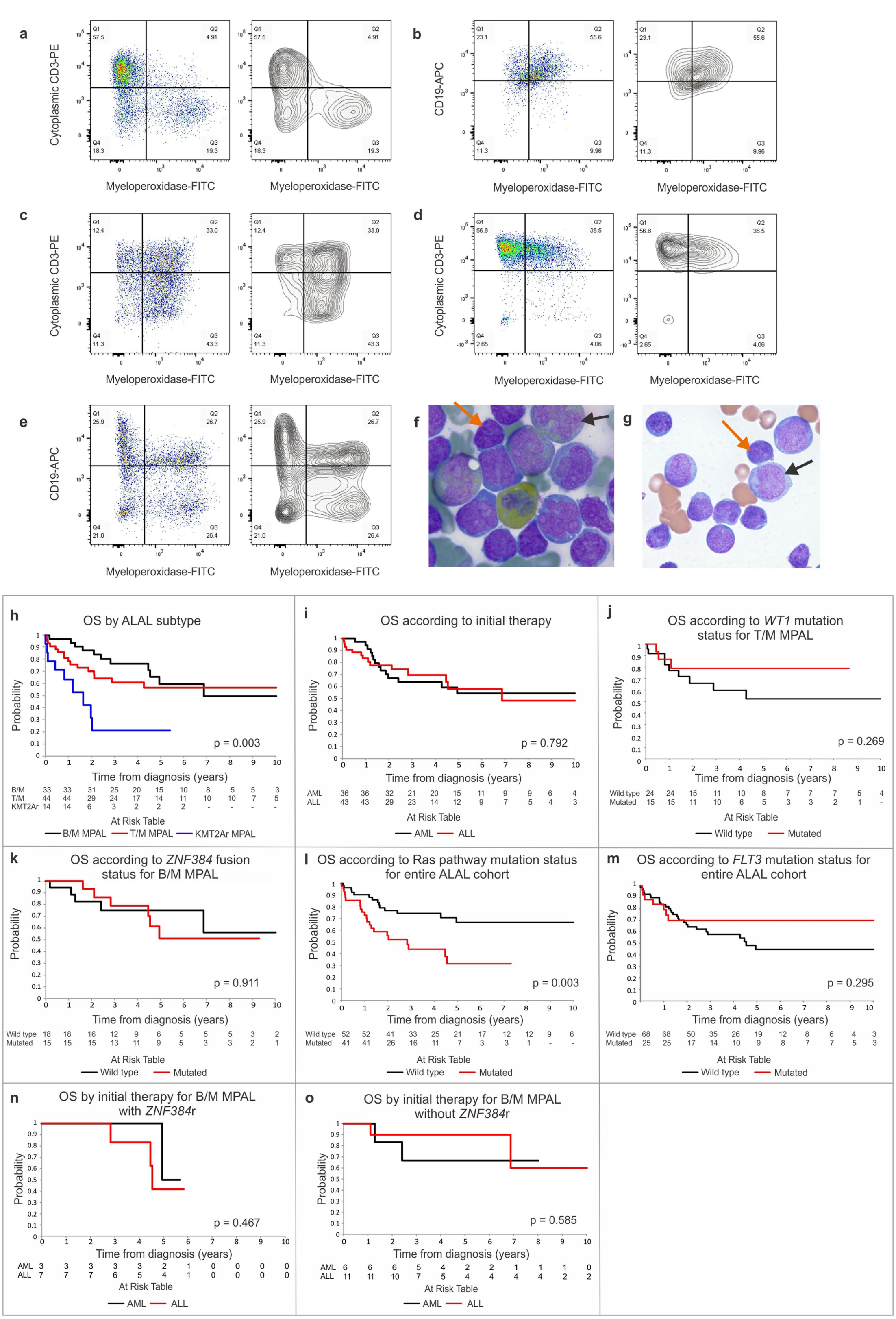
Mixed Phenotype Acute Leukemia - Leukemia blasts express specific antigens from multiple leukocyte lineages
- ***ZNF384* rearranged** - Leukemia blasts express antigens from multiple lineages AND the presence of a *ZNF384* rearrangement
- **Ph-like** - Leukemia blasts express antigens from multiple lineages AND the presence of a Ph-like gene expression profile
- ***KMT2A* rearranged** - Leukemia blasts express antigens from multiple lineages AND the presence of a *KMT2A* rearrangement
- ***BCR-ABL* positive** - Leukemia blasts express antigens from multiple lineages AND the presence of a *BCR-ABL* fusion
- **T/myeloid, with *WT1* mutations** - Leukemia blasts express both T-lymphoid and myeloid antigens AND the presence of a *WT1* mutation
- **B/myeloid, NOS** - Leukemia blasts express both B-lymphoid and myeloid antigens without a recurrent genetic abnormality
- **T/myeloid, NOS** - Leukemia blasts express both T-lymphoid and myeloid antigens without a recurrent genetic abnormality
- **Not otherwise specified** - Leukemia blasts express both B and T-lymphoid antigens OR T-lymphoid, B-lymphoid, and myeloid antigens without a recurrent genetic abnormality

Acute Undifferentiated Leukemia - Leukemia blasts do not express any lineage defining antigens
- **Ph-like** - Leukemia blasts do not express any lineage defining antigens AND the presence of a Ph-like gene expression profile
- ***KMT2A* rearranged** - Leukemia blasts do not express any lineage defining antigens AND the presence of a *KMT2A* rearrangement
- **Not otherwise specified** - Leukemia blasts do not express any lineage defining antigens AND there is no recurrent genetic abnormality

d

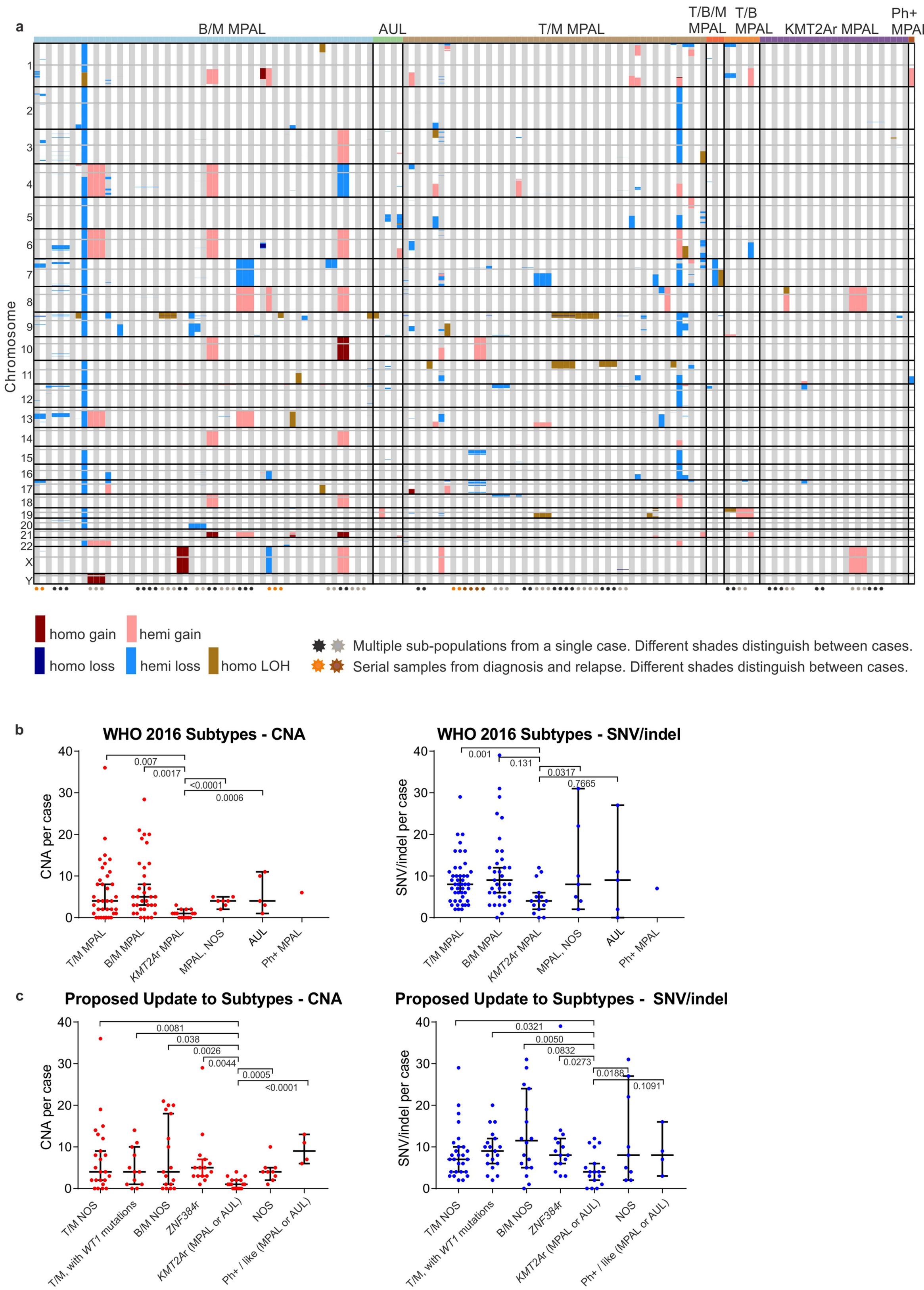


Extended Data Fig. 1 | Criteria for diagnosis of ALAL. a, Subtypes of ALAL according to the WHO 2008 criteria and consistent with minor revisions of WHO 2016 criteria[6]. **b**, Antigen requirements for lineage assignment for MPAL according to WHO 2008 criteria. The 2016 revisions to the WHO classification for ALAL did not change the above categories or requirements. Rather, the revision emphasized that care should be taken before making a diagnosis of B/M MPAL when low-intensity myeloperoxidase is the only myeloid-associated feature. Additionally, the revision emphasized that in cases in which it is possible to resolve two distinct blast populations, it is not necessary that the specific markers be present, but only that each population would meet the criteria for B, T, or myeloid leukaemia[64]. **c**, Proposed update to WHO ALAL subtypes incorporating critical newer genomic information (new subtypes in red). **d**, Flow chart of ALAL cohort showing reasons for exclusion and initial diagnosis in cases for which initial ALAL diagnosis occurred at relapse.



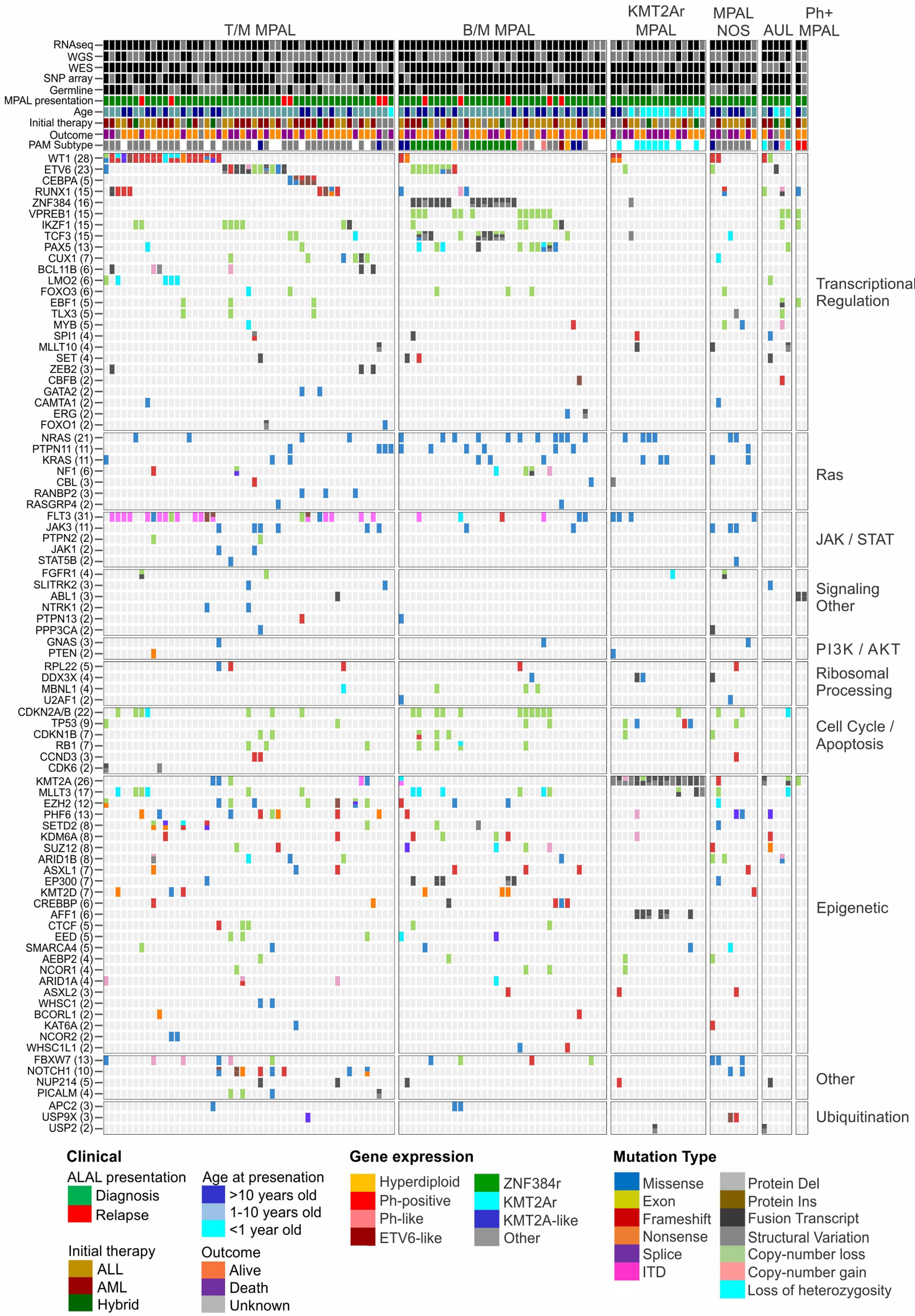
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Illustrative immunophenotype and overall survival. **a**–**e**, Representative flow cytometry pseudocolour dot plots and contour plots for five different MPAL cases gated on blast area from CD45 and side scatter area (SSC-A). There are a wide variety of immunophenotypic patterns, including classic bilineal phenotype (**a**), classic biphenotypic case (**b**), myeloid predominance (**c**), lymphoid predominance (**d**) and complex phenotype with more than two immunophenotypic clones (**e**). **f**, **g**, Morphology of cells from two patients with MPAL showing both lymphoid (orange arrow) and myeloid (black arrow) morphology. **f**, Bone marrow aspirate stained with myeloperoxidase from a patient with T/M MPAL showing multiple blasts with moderate MPO positivity along with one normal granulocyte. **g**, Peripheral blood haematoxylin and eosin stain from a patient with B/M MPAL. **h**–**o**, Kaplan–Meier survival curves with overall survival (OS) distributions of patients whose initial diagnosis was MPAL or AUL compared using log-rank tests. At risk numbers for each analysis are provided in the figures. Outcome associations were analysed with the log-rank test. Overall survival according to WHO 2016 subtype (**h**), initial therapy (**i**), *WT1* status within the T/M MPAL cohort (**j**), *ZNF384* status within the B/M MPAL cohort (**k**), Ras pathway alteration within the entire cohort (**l**) and *FLT3* alteration within the entire cohort (**m**). **n**, Overall survival according to initial therapy for patients with B/M MPAL with *ZNF384*r. **o**, Overall survival according to initial therapy for patients with B/M MPAL without *ZNF384*r. Patients included in this cohort were collected from a range of treatment eras, treatment locations, treatment regimens, and include a range of ages and genomic subtype, limiting the conclusions that can be drawn from these analyses.

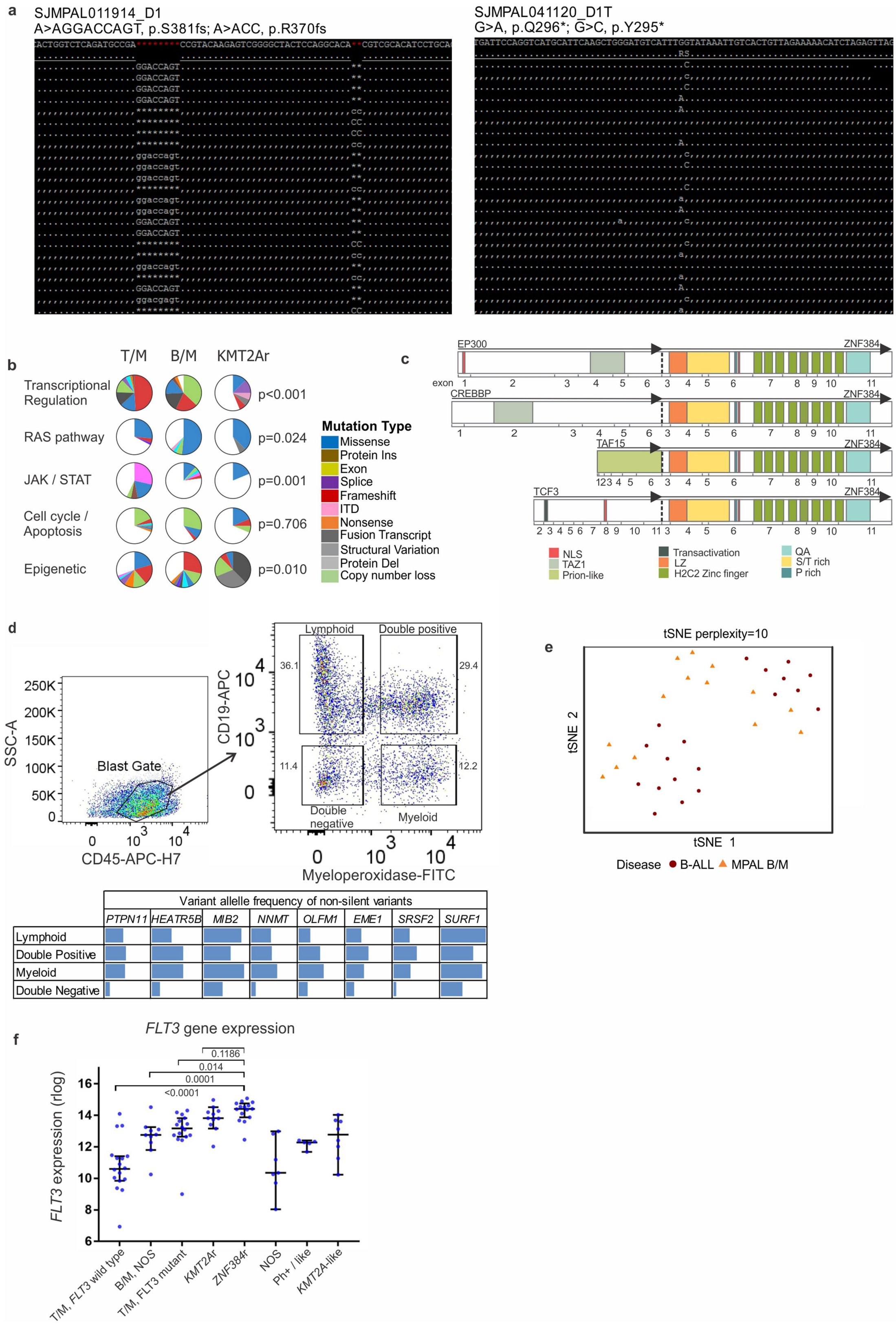


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Copy number alterations and mutation burden in ALAL. a, Map showing spectrum of CNAs, visually recapitulating the data shown in Supplementary Table 10. Twenty-seven patients had SNP arrays for multiple subpopulations, annotated by stars. **b**, CNA and non-silent SNVs or indels in ALAL subtypes according the WHO 2016 classification. (CNA, T/M MPAL $n=36$, B/M MPAL $n=34$, *KMT2A*r MPAL $n=15$, MPAL NOS $n=7$, AUL $n=5$, Ph+ MPAL $n=1$; SNV/indel, T/M MPAL $n=46$, B/M MPAL $n=35$, *KMT2A*r MPAL $n=15$, MPAL NOS $n=7$, AUL $n=5$, Ph+ MPAL $n=1$.) Patients with *KMT2A*r MPAL have a lower mutation burden than those with T/M MPAL or B/M MPAL. **c**, CNAs and non-silent SNVs or indels in our proposed updated classification system. (CNA, T/M MPAL NOS $n=24$, T/M MPAL with *WT1* alteration $n=12$, B/M MPAL NOS $n=17$, B/M MPAL with *ZNF384*r $n=15$, *KMT2A*r MPAL/AUL $n=17$, MPAL/AUL NOS $n=9$, Ph+/Ph-like MPAL/AUL $n=4$; SNV/indel, T/M MPAL NOS $n=27$, T/M MPAL with *WT1* alteration, $n=19$, B/M MPAL NOS $n=18$, B/M MPAL with *ZNF384*r $n=15$, *KMT2A*r MPAL/AUL $n=17$, MPAL/AUL NOS $n=9$, Ph+/Ph-like MPAL/AUL $n=4$.) Data shown as median ± 95% confidence interval. Comparisons assessed by two-sided unpaired *t*-test. One data point is outside the SNV/indel graph for the B/M NOS subtype (1 patient with 167 SNV/indels). SNV/indels per case are shown for cases with completed DNA sequencing analyses.

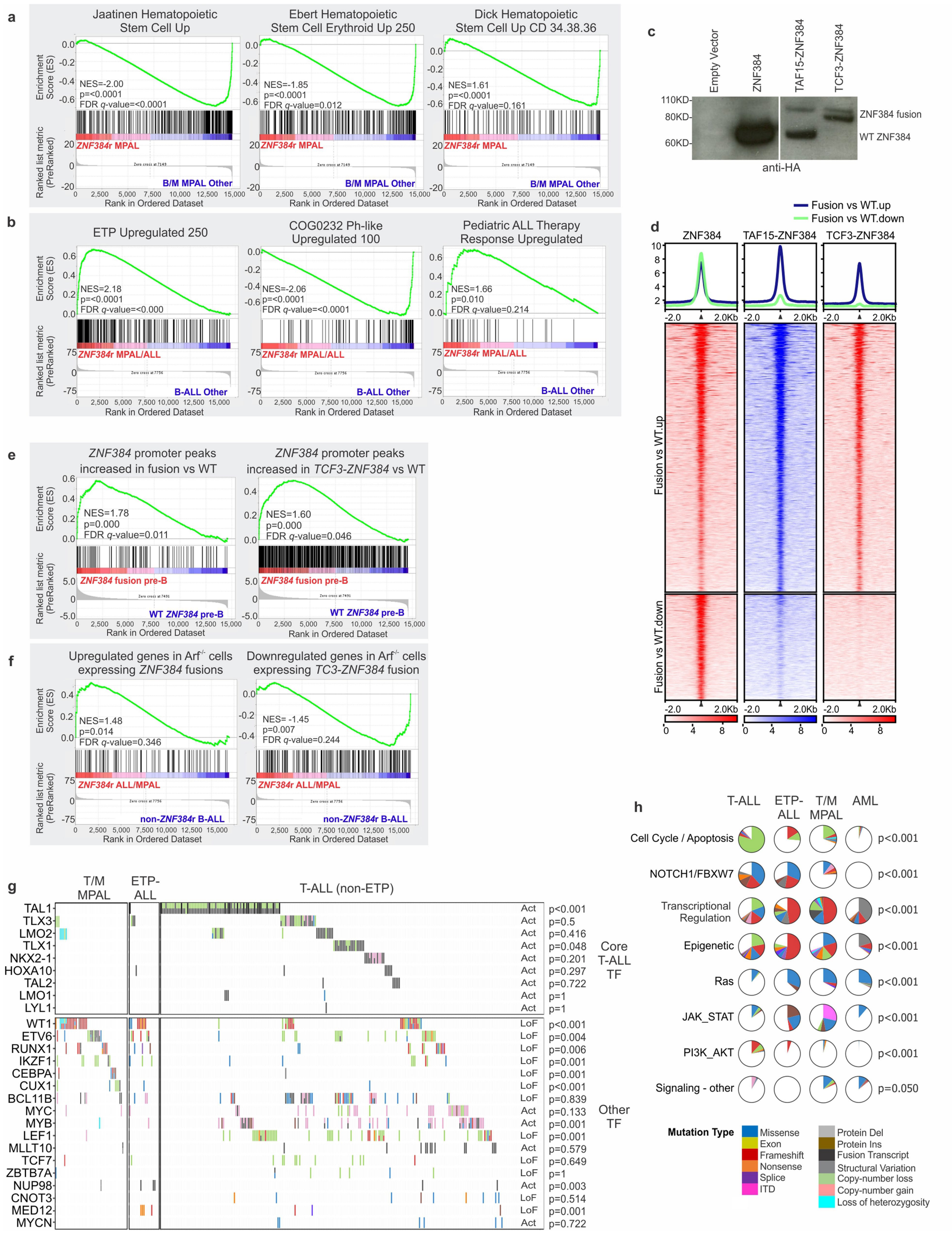


Extended Data Fig. 4 | Complete ALAL mutation oncoprint. Mutation spectrum of ALAL.



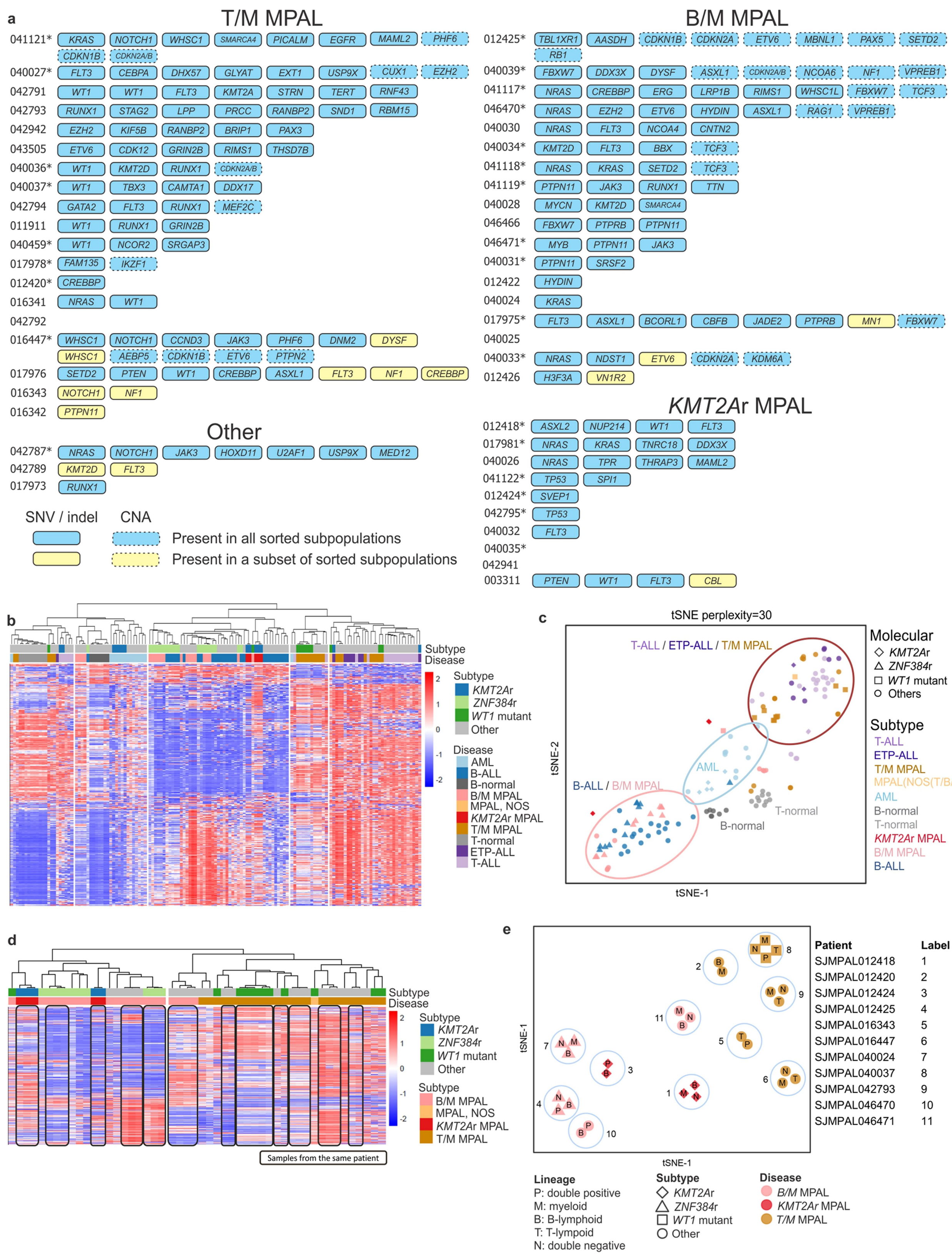
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Features of MPAL genomic analysis. **a**, *WT1* alterations were observed in 28 patients, commonly as frameshift mutations (31/47 mutations) in exon 7 (29/47 mutations) and were frequently biallelic. In 16 patients, two clonal alterations were detected, and in 9 patients the locations of the alteration were encompassed by the same sequencing read, providing definitive demonstration that the mutations were *in trans*. Additionally, one patient (SJMPAL043773) had a frameshift mutation and copy number loss of the second allele, while another had a frameshift mutation with copy-neutral loss of heterozygosity (SJMPAL040036). Data are shown for two representative patients with MPAL, showing double-hit mutations on *WT1*. The read alignment view was generated by Samtools[24]. The reference human genome is on the first row and sequence reads are aligned below, with matched nucleotides as dots (forward strand match) and commas (reverse strand match) and mismatched ones showing the differences. Alignment gaps are shown as asterisks. Adjacent mutations are shown on different sequence reads, indicating that the mutations are on different alleles. **b**, Frequency of alteration by pathway analysis and MPAL subtype. The similarity of somatic alteration prevalence in different leukaemia subtypes was evaluated by two-sided Fisher's exact test ($n = 100$ biologically independent cases). See also Supplementary Tables 12, 13 for numbers and *P* values for each gene and pathway. **c**, Schematic representation of *ZNF384*r observed in B/M MPAL. NLS, nuclear localization signal; TAZ1, transcriptional adaptor zinc-binding; LZ, leucine rich domain; QA, glycine/alanine repeat. **d**, FACS schema in a representative case with a *ZNF384*r, and VAF of SNVs/indels present in the respective sorted subpopulations, demonstrating genomic similarity of the sorted populations. **e**, *t*-SNE plot of RNA-seq gene expression of all patients with *ZNF384r* show no clear segregation of B/M MPAL and B-ALL cases. **f**, *FLT3* gene expression in subtypes of ALAL showing that patients with *ZNF384*r B/M MPAL have high levels of *FLT3* expression. As in patients with *KMT2A*r, this occurs in the absence of *FLT3* alteration in most cases. By contrast, high levels of *FLT3* expression in T/M MPAL appears to be driven by *FLT3* alterations. Data shown as median ± 95% confidence interval. Comparisons assessed by unpaired *t*-test, two sided. T/M MPAL *FLT3* wild type $n = 18$, B/M MPAL NOS $n = 10$, T/M MPAL with *FLT3* alteration $n = 16$, B/M MPAL NOS $n = 17$, B/M MPAL with *ZNF384*r $n = 15$, *KMT2A*r MPAL/AUL $n = 11$, MPAL/AUL NOS $n = 7$, Ph+/Ph-like MPAL/AUL $n = 5$, KMT2A-like MPAL/AUL $n = 8$.



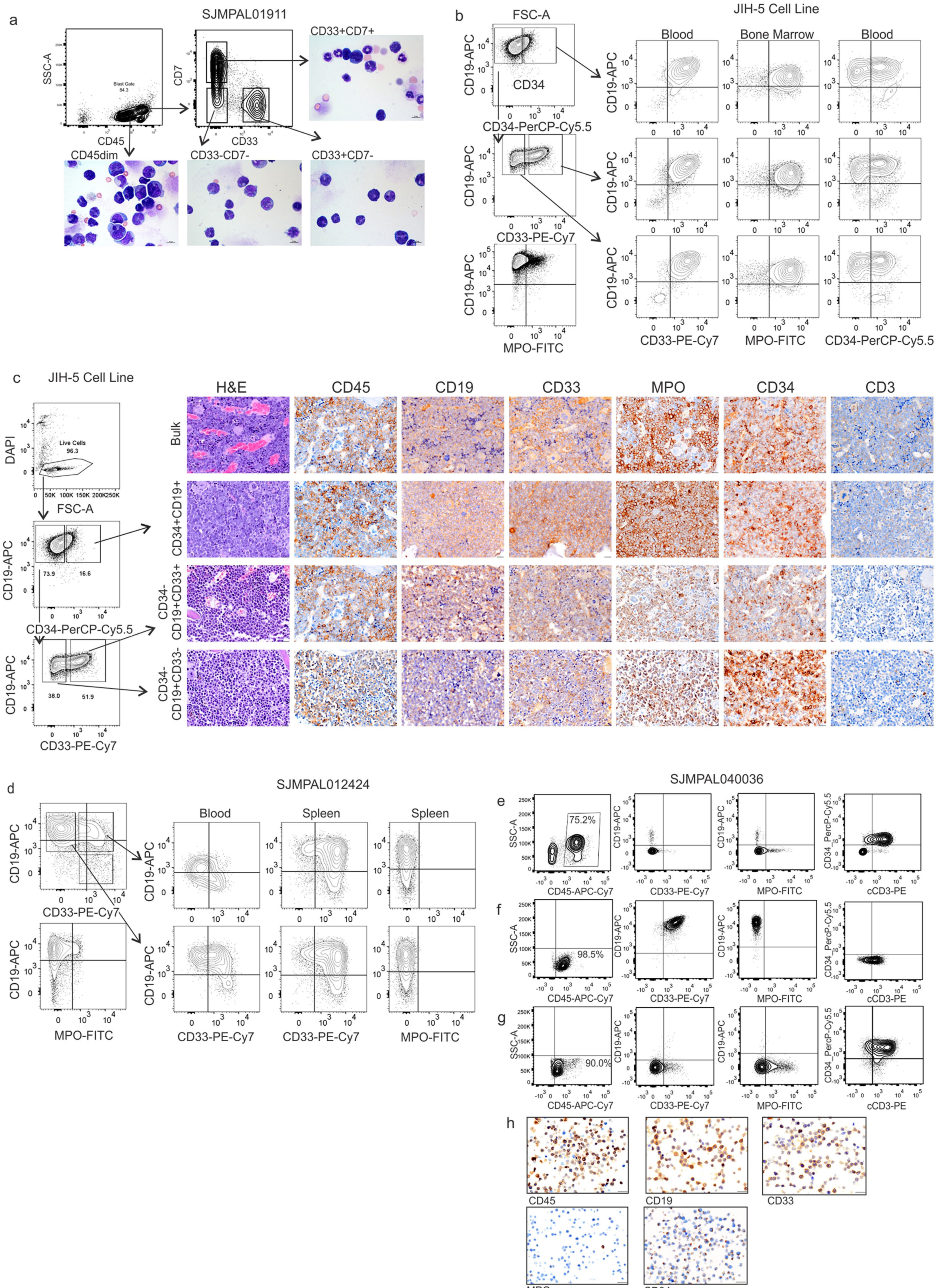
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | *ZNF384*r leukaemia analysis and T/M MPAL mutation comparisons. **a**, GSEA of *ZNF384*r B/M MPAL versus non-*ZNF384*r B/M MPAL. HSC gene sets are negatively enriched, supporting the proposed update to MPAL subtypes in which *ZNF384*r leukaemia has distinct biology compared with other B/M MPAL cases[20,65,66]. **b**, GSEA of all *ZNF384*r cases versus other B-ALL cases indicates immaturity of this subtype compared to B-ALL, with positive enrichment for genes upregulated in ETP-ALL (a stem cell leukaemia), and negative enrichment for genes upregulated in Ph-like ALL in other B-ALL cases. *ZNF384*r acute leukaemia is also enriched for genes upregulated in patients with detectable minimal residual disease at end of induction[10,51,67]. **c**, Western blot analysis to validate expression of *ZNF384*, *TAF15–ZNF384*, and *TCF3–ZNF384* in transduced $Arf^{-/-}$ pre-B cells. Proteins contain an HA epitope tag and are detected by anti-HA antibody. **d**, Heat map showing the ChIP–seq signal, centred on *ZNF384* peaks, of wild-type (WT) *ZNF384* compared to *TAF15–ZNF384* and *TCF3–ZNF384*. Middle, peaks with increased binding of fusion proteins compared to wild-type proteins. Bottom, peaks with decreased binding of the fusion proteins compared to wild-type proteins. **e**, GSEA showing enrichment of genes whose promoters exhibit increased binding by ZNF384 fusions in the GEP of ZNF384r versus wild-type pre-B cells. **f**, GSEA showing similarity of the GEP of mouse pre-B cells expressing ZNF384r to the GEP of human *ZNF384*r leukaemia cells, supporting the notion that perturbation of ZNF384 binding contributes to deregulated gene expression in human *ZNF384*r leukaemia. **g**, Oncoprint of mutations in transcription factor genes across T/M MPAL ($n=49$), ETP-ALL ($n=19$) and T-ALL (other) ($n=245$), showing lack of *TAL1* alterations in T/M MPAL and few core T-ALL transcription factor alterations in T/M MPAL or ETP-ALL. The association of leukaemia subtype with individual transcription factor alterations was evaluated using two-sided Fisher exact test. Act, activating mutation; LoF, loss-of-function mutation. **h**, Gene pathway analyses showing similarity of ETP-ALL and T/M MPAL, specifically in frequency of mutations in pathways regulating cell cycle or apoptosis, transcriptional regulation, and signalling pathways. The similarity of somatic alteration prevalence in different leukaemia subtypes was evaluated by two sided Fisher's exact tests in these four subtypes (T/M MPAL $n=49$, ETP-ALL $n=19$, non-ETP T-ALL $n=245$, AML $n=197$).



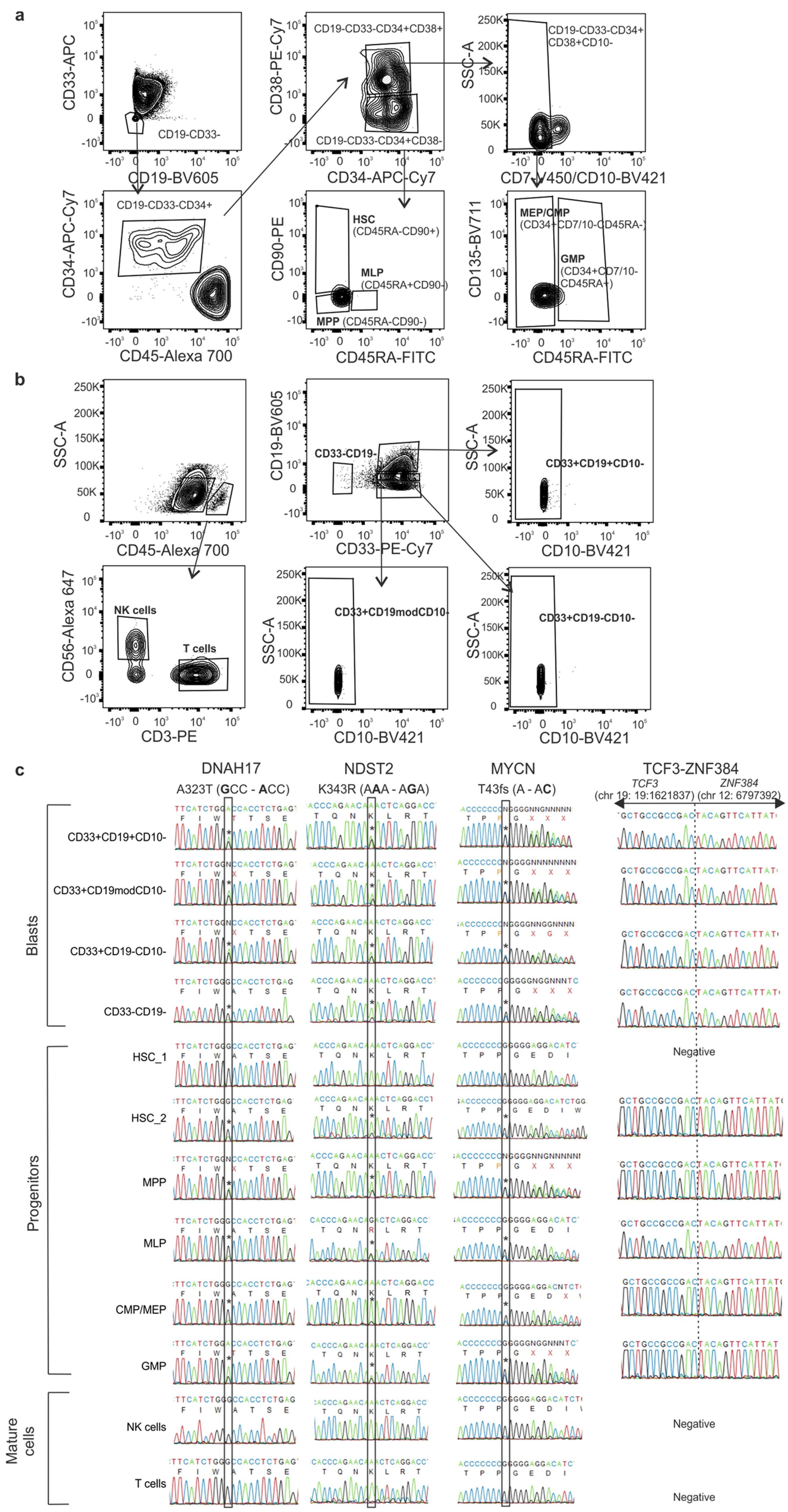
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | MPAL subpopulation analysis and methylation analysis. a, Results of genomic analysis of the 50 patients with sorted subpopulations with WGS or WES results. Listed here are all genes with mutations that were either recurrent in the ALAL cohort or were in known cancer census genes[68]. *CNA results also available for sorted subpopulations in these cases. **b**–**d**, Methylation analysis of MPAL, comparison with acute leukaemia and normal lymphocytes. The top 5,000 probes with highest mean absolute deviation were used to assess the clustering through a 2D *t*-SNE plot and heat map with Pearson correlation clustering. See Supplementary Table 37 for sample details. **b**, Heat map of all samples used for methylation analysis showing the general alignment of samples by leukaemia phenotype with B/M cases clustering with B-ALL, T/M MPAL, ETP-ALL cases together, and AML cases clustering separately. **c**, *t*-SNE analysis of the same samples as in the top heat map, showing general alignment by leukaemia phenotype with B/M cases clustering with B-ALL, T/M MPAL, ETP-ALL cases together, and AML cases clustering separately. **d**, Heat map of all MPAL cases, again showing some clustering by phenotype between B/M and T/M cases. Subpopulations sorted by distinct immunophenotype in MPAL cases clustered tightly with samples from the same patient, rather than with samples with similar phenotype from a different patient. **e**, Methylation analysis of sorted subpopulations from 11 patients with MPAL, demonstrating that methylation profiles cluster by patient and not by immunophenotype lineage.



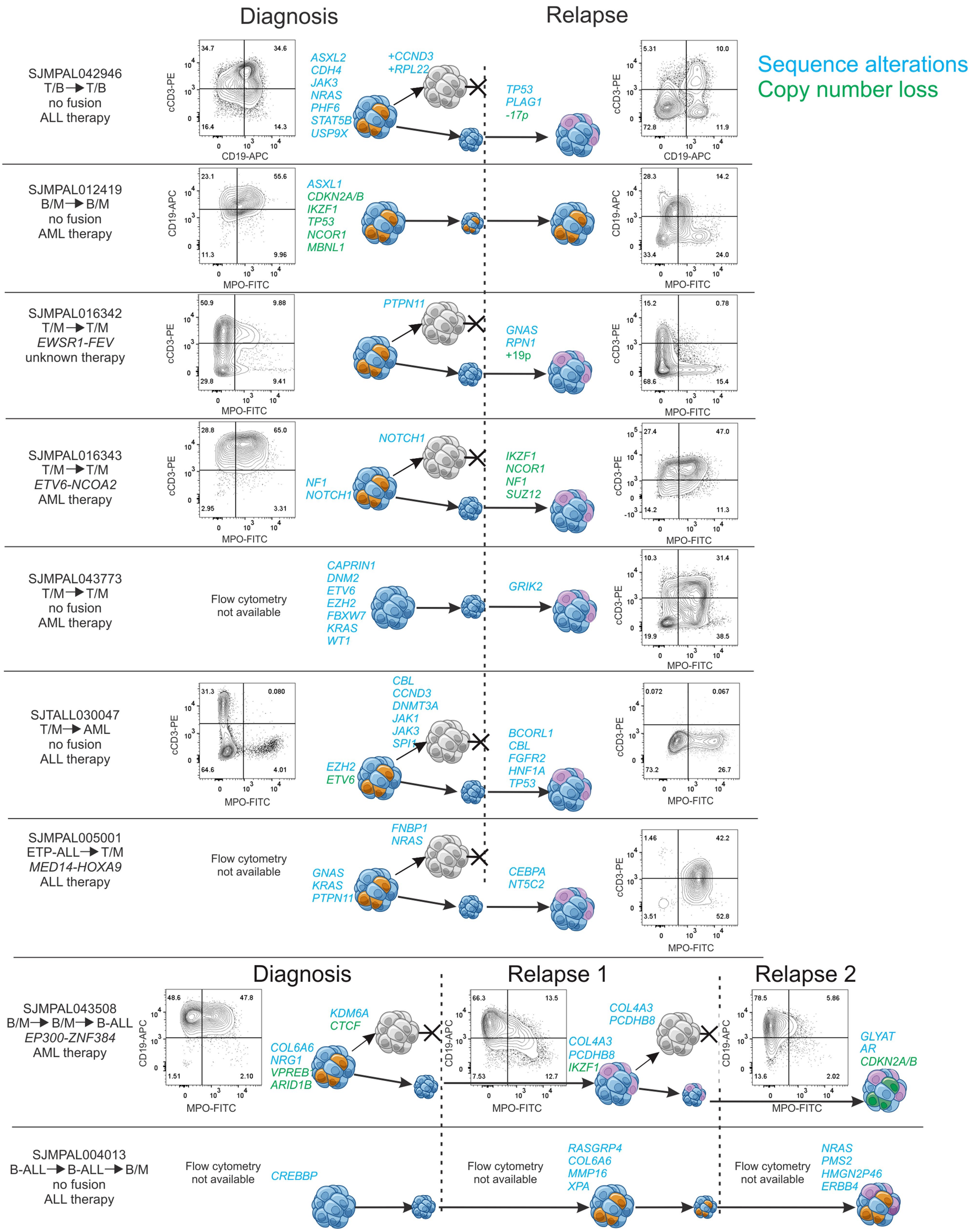
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Xenograft analysis. **a**, Flow cytometry analysis of bulk leukaemic cells from patient SJMPAL011911 before sorting, and cytospins from bone marrow samples from representative primary recipient mice transplanted with different leukaemia subpopulations or bulk, confirming the presence of leukaemic blasts from each engrafted population. Scale bars, 10 μm. **b**, Phenotypic subpopulations from JIH-5 cells in the first column were sorted and injected into NSG-SGM3 mice. Remaining plots show the immunophenotypes of engrafted leukaemia propagated from each sorted subpopulation, demonstrating recapitulation of biphenotypic leukaemia from each. **c**, Flow cytometry analysis of bulk JIH-5 cells prior to sorting (left) and haematoxylin and eosin staining and IHC labelling for human CD45, CD19, CD33, MPO, CD34 and CD3 in sternum samples from representative primary recipient mice transplanted with different leukaemia subpopulations or bulk. Scale bars, 20 μm. **d**, Phenotypic subpopulations from patient SJMPAL012424 were sorted (left) and injected into irradiated NSG-SGM3 mice. Remaining plots show the immunophenotypes of engrafted leukaemia from each starting subpopulation, demonstrating recapitulation of mixed phenotype leukaemia from two sorted subpopulations. **e**, Flow cytometry analyses of bone marrow cells from an engrafted primary mouse transplanted with leukaemia cells from a patient with T/M MPAL (SJMPAL040036). **f**, **g**, Flow cytometry analyses of representative engrafted secondary recipient mice transplanted with leukaemia cells from the mouse in **e** showing lineage plasticity with mice developing an emerging $CD19^{+}CD33^{+}$ population (**f**) and other mice recapitulating the immunophenotype in the primary recipient (**g**). **h**, IHC labelling for human CD45, CD19, CD33, MPO and CD34 from harvested and fixed spleen cells from a representative secondary recipient mouse showing high expression of CD19 and CD33 and thus confirming the leukaemic lineage plasticity. Scale bars, 20 μm.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Haematopoietic progenitor cell analysis. **a**, Progenitor cell sorting scheme for diagnosis sample from patient SJMPAL040028. Progenitor populations were all gated on $CD19^{-}CD33^{-}CD34^{+}$ and sorted into HSC ($CD38^{-}CD34^{+}CD90^{+}CD45RA^{-}$; 2 replicates: HSC_1 and HSC_2); MPP ($CD38^{-}CD34^{+}CD90^{-}CD45RA^{-}$); MLP ($CD38^{-}CD34^{+}CD45RA^{+}$); megakaryocyte erythroid progenitors/common myeloid progenitors (MEP/CMP; $CD38^{+}CD34^{+}CD7^{-}CD10^{-}CD45RA^{-}$); and granulocyte monocyte progenitor (GMP; $CD38^{+}CD34^{+}CD7^{-}CD10^{-}CD45RA^{+}$) populations. **b**, Blast cell sorting scheme for diagnosis sample from patient SJMPAL040028. Cells were gated on $CD45^{dim}$ and sorted into four different immunophenotypic populations ($CD33^{+}CD19^{+}CD10^{-}$; $CD33^{+}CD19^{mod}CD10^{-}$; $CD33^{+}CD19^{-}CD10^{-}$; and $CD33^{-}CD19^{-}$). **c**, Sanger sequencing electropherograms for the mutational status of *DNAH17*, *NDST2* and *MYCN* and for the fusion *TCF3–ZNF384* in isolated progenitor and blast populations from patient SJMPAL040028 at diagnosis. The identification of somatic missense mutations and *TCF3–ZNF384* fusion in early haematopoietic progenitors indicate that the ambiguous phenotype of MPAL is the result of the acquisition of alterations within an immature haematopoietic progenitor cells.



Extended Data Fig. 10 | Phenotypic and genotypic evolution from diagnosis to relapse. Patients for whom diagnosis and relapse pairs with matching non-tumour controls are available show recapitulation of the diagnostic multilineage phenotype in some cases and phenotype plasticity in others. The first column shows the case ID, the leukaemia subtype at diagnosis and then subsequent relapse, the in-frame fusion if present, and initial therapy received by the patient. Flow plots are shown of cells gated on CD45^{dim} versus SSC-A^{low}. The diagram depicts the inferred clonal evolution based on WES and/or WGS and SNP array data (where available). Mutated genes (either recurrent in ALAL cohort or known cancer census genes[68]) are listed. The genes beside the initial diagnostic cell cluster remained present at relapse. The grey cells represent clones that were extinguished with therapy. The genes in the relapse column represent mutations that were gained at relapse.

https://doi.org/10.1038/s41586-018-0569-1

Transcriptional recording by CRISPR spacer acquisition from RNA

Florian Schmidt[1], Mariia Y. Cherepkova[1] & Randall J. Platt[1,2]*

The ability to record transcriptional events within a cell over time would help to elucidate how molecular events give rise to complex cellular behaviours and states. However, current molecular recording technologies capture only a small set of defined stimuli. Here we use CRISPR spacer acquisition to capture and convert intracellular RNAs into DNA, enabling DNA-based storage of transcriptional information. In *Escherichia coli*, we show that defined stimuli, such as an RNA virus or arbitrary sequences, as well as complex stimuli, such as oxidative stress, result in quantifiable transcriptional records that are stored within a population of cells. We demonstrate that the transcriptional records enable us to classify and describe complex cellular behaviours and to identify the precise genes that orchestrate differential cellular responses. In the future, CRISPR spacer acquisition-mediated recording of RNA followed by deep sequencing (Record-seq) could be used to reconstruct transcriptional histories that describe complex cell behaviours or pathological states.

A central challenge in biology is to understand how the molecular components of a cell function and integrate to enable complex cell behaviours. This challenge has fuelled the creation of increasingly sophisticated technologies that facilitate detailed intracellular observations at the levels of DNA, RNA, protein, and metabolites[1]. In particular, RNA sequencing technologies enable researchers to quantify transcriptomes within multiple or single cells, revealing the molecular signatures of cell behaviours, states, and types with unprecedented detail[2,3]. Despite the power of these technologies, they require destructive methods and therefore observations are limited to a few snapshots in time or to select asynchronous cellular processes. One provocative solution to this is to introduce synthetic memory devices[4] within cells that enable the encoding, storage, and retrieval of transcriptional information.

The bacterial adaptive immune system CRISPR–Cas embodies the ideal molecular recorder. Molecular memories of plasmid or viral infections are stored within CRISPR arrays in the form of short nucleic acid segments (spacers) separated by direct repeats. New memories are acquired via the action of a complex of Cas1 and Cas2, which integrates new spacers ahead of old spacers within the CRISPR array, thereby providing a temporal memory of molecular events[5–14]. The prototype type I-E CRISPR acquisition system from *E. coli* was recently leveraged to store arbitrary information[15,16] and quantifiable records of defined stimuli within bacterial populations[17]. These systems elegantly demonstrate the potential of using CRISPR spacer acquisition as a molecular recorder, but they are currently limited by the need to electroporate chemically synthesized nucleotides[15,16] or, analogous to previous technologies, the availability of inducible promoters[15–25]. Moreover, these systems acquire spacers derived from DNA but not RNA, and therefore do not globally reflect the transcriptional history of a cell. Although one naturally occurring fusion protein between Cas1 and a reverse transcriptase (RT) domain (RT–Cas1) was recently experimentally validated to acquire spacers directly from RNA, it did not maintain this function heterologously in *E. coli*[26].

We hypothesized that direct CRISPR spacer acquisition from RNA could be leveraged to store transcriptional records in CRISPR arrays within living cells (Fig. 1a, b). Therefore, we characterized several orthologous RT–Cas1-containing CRISPR–Cas systems and found that one from *Fusicatenibacter saccharivorans* could acquire RNA spacers heterologously in *E. coli*. Leveraging *F. saccharivorans* RT–Cas1 and Cas2 (*Fs*RT–Cas1–Cas2), we developed Record–seq, a method that enables transcriptome-scale molecular recordings into populations of cells. Transcriptional events are recorded according to RNA abundance, stored in CRISPR arrays within DNA, and can be used to describe both continuous and transient complex cellular behaviours.

CRISPR spacer acquisition by *Fs*RT–Cas1–Cas2

We set out to identify an RT–Cas1–Cas2 CRISPR acquisition complex that could acquire spacers directly from RNA upon heterologous expression in *E. coli*. We identified 121 RT–Cas1 orthologues (Supplementary Table 1), and selected 14 of these for functional characterization (Extended Data Fig. 1a, b). We overexpressed corresponding RT–Cas1 and Cas2 proteins from a plasmid that also contained their predicted CRISPR array (Extended Data Fig. 1a, Supplementary Table 2). Using a previously established spacer acquisition assay[27], we found that only one of the orthologues tested (*F. saccharivorans*) actively acquired new spacers (Extended Data Fig. 1c). The endogenous *F. saccharivorans* locus contains two CRISPR arrays and we observed that novel spacers derived from the overexpression plasmid as well as the *E. coli* genome were acquired into either array (Extended Data Fig. 1c–e).

Selective amplification of expanded CRISPR arrays

Using the previously established spacer acquisition assay[27], we obtained approximately 1,300 newly acquired spacers per 1 million deep sequencing reads for *Fs*RT–Cas1–Cas2 (Extended Data Fig. 1c). To improve detection of novel spacers, we developed 'selective amplification of expanded CRISPR arrays' (SENECA), a method for selective amplification of CRISPR arrays that acquired new spacers (Fig. 2a, Extended Data Fig. 2a). A typical SENECA-assisted Record–seq experiment uses an input of about 180 ng plasmid DNA extracted from an overnight culture of *E. coli* overexpressing *Fs*RT–Cas1–Cas2, and yields 950,000 total spacers aligning to the plasmid or host genome for every 1 million sequencing reads (Fig. 2a, Extended Data Fig. 2b–e, Supplementary Notes). This represents an improvement of several thousand-fold compared to recent reports[26,28]. Using Record–seq,

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland. [2]Department of Chemistry, University of Basel, Basel, Switzerland. *e-mail: rplatt@ethz.ch

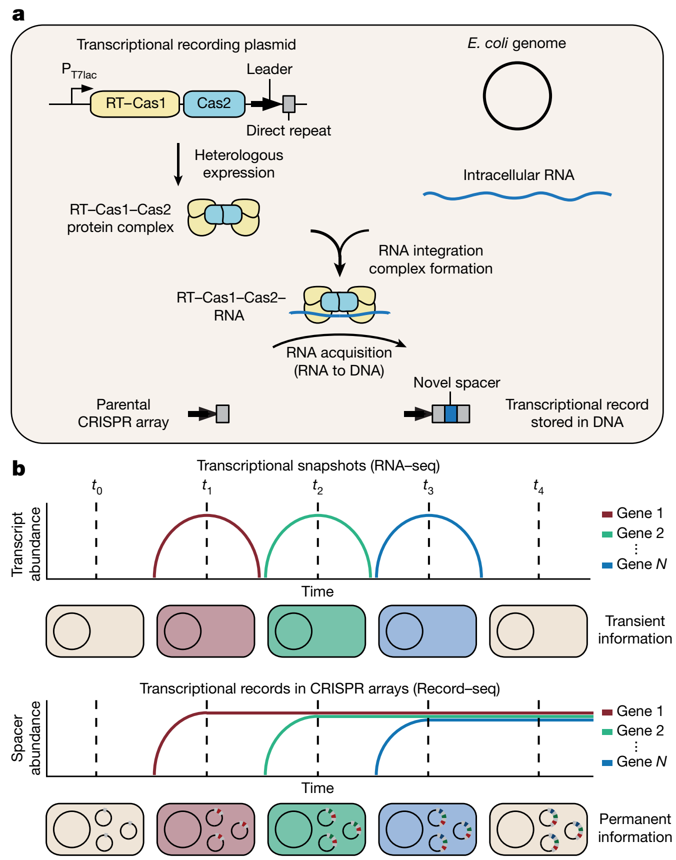


Fig. 1 | Transcriptional recording by CRISPR spacer acquisition from RNA. a, Expression of RT–Cas1–Cas2 leads to the acquisition of intracellular RNAs, providing a molecular memory of transcriptional events stored within DNA. **b**, Comparison of RNA-seq and Record–seq. RNA-seq captures the transcriptome of a population of cells at a single point in time, providing a transient snapshot of cellular events. By contrast, Record–seq permanently stores information about prior transcriptional events in a CRISPR array, providing a molecular record that can be used to reconstruct transcriptional events that occurred over time.

we readily demonstrated in vivo activity of *Fs*RT–Cas1–Cas2 in various *E. coli* strains and throughout growth phases (Extended Data Fig. 2b–g).

We then used Record–seq to rescreen our initial selection of RT–Cas1 orthologues (Extended Data Fig. 1b). Furthermore, we included all potential CRISPR arrays present in their endogenous loci in both possible directions in order to overcome the challenges associated with predicting these a priori (Extended Data Fig. 3a). Owing to the improved sensitivity of Record–seq compared to the classic readout, we readily detected newly acquired spacers for the majority of orthologues upon RT–Cas1–Cas2 expression (Extended Data Fig. 3b). Only a few orthologues exhibited a preferred directionality of the CRISPR array (that is, specificity for an upstream leader sequence). Consistent with the classic readout, *Fs*RT–Cas1–Cas2 outperformed all other orthologues in terms of spacer acquisition efficiency and was chosen for further characterization. The concepts on which Record–seq is based could also be used to characterize spacer acquisition in other CRISPR–Cas systems that have been intractable owing to low spacer acquisition efficiencies.

Characteristics of *Fs*RT–Cas1–Cas2 spacer acquisition

To better understand the properties of *Fs*RT–Cas1–Cas2, we extensively characterized newly acquired spacers by performing Record–seq on populations of *E. coli* overexpressing *Fs*RT–Cas1–Cas2 (Fig. 2a). We observed that genome-aligning spacers were preferentially acquired with a specific 'antisense' orientation, whereby spacers were complementary to the originating RNA (Fig. 2b, c). The median spacer length was 39 bp, with a distribution biased towards longer lengths (Fig. 2d). The median GC content was 36%, showing a strong bias towards AT-rich spacers (Fig. 2e). In line with previously described type III CRISPR systems[29], we did not find a sequence preference within or adjacent to newly adapted spacers acquired from either plasmid (Extended Data Fig. 4a) or genome (Fig. 2f), indicating that the *Fs*RT–Cas1–Cas2 complex has no protospacer adjacent motif (PAM). While observing spacer alignments to the *E. coli* genome, we noted that many coverage peaks were located near the termini of genes (Fig. 2b). Consistent with this observation, we found that at the genome-wide level, most spacers were derived from the 5′, and to a lesser extent, 3′ ends of genes (Fig. 2g). This finding raised the possibility that the apparent bias towards AT-rich spacers might be caused by the AT-richness of RNA ends in *E. coli*, but the bias towards AT-rich spacers persisted when considering only spacers derived from within the gene body (Extended Data Fig. 4b). We directly compared SENECA with the classic spacer readout to determine whether SENECA introduced additional biases, but found no major differences (Extended Data Fig. 4c–h). Together, these results reflect a process by which *Fs*RT–Cas1–Cas2 selects AT-rich spacer-based sequences related to the beginning or end of a gene, such as the ends of an RNA molecule.

*Fs*RT–Cas1–Cas2 acquires spacers directly from RNA

To determine whether *Fs*RT–Cas1–Cas2 acquires spacers directly from RNA, we used a self-splicing *td* group I intron[30–32]. This intron is a functional ribozyme that catalyses its own excision from the pre-mRNA, resulting in a characteristic splice junction that is not present at the DNA level. We constructed three intron-interrupted constructs based on genes that were highly sampled by spacers, namely *cspA*, *rpoS* and *argR* (Fig. 3a). Upon expression of these constructs followed by Record–seq, we observed unique spacers spanning the splice junctions (Fig. 3a, b). To exclude the possibility that splice junction-containing spacers were acquired from extended complementary DNA copies generated through nonspecific reverse transcriptase activity in *E. coli*, we performed targeted deep sequencing on genomic DNA extracted from cultures expressing *td* intron constructs (Extended Data Fig. 5a), and found that the splice junction was absent at the DNA level (Extended Data Fig. 5a, b). Notably, these results do not exclude the possibility of spacer acquisition from DNA (Supplementary Notes). Thus, *Fs*RT–Cas1–Cas2 facilitates CRISPR spacer acquisition from RNA heterologously in *E. coli*.

To further validate this finding, we used the Enterobacteria phage MS2. MS2 phages exist as both sense and antisense single-stranded RNAs during their lifecycle but have no DNA intermediates. Given that MS2 phages require the F pilus for cell entry, which is missing in *E. coli* BL21(DE3) cells, we turned to the *E. coli* K12 strain NovaBlue(DE3). Upon infection of *Fs*RT–Cas1–Cas2-expressing cells with MS2 phage, we could readily observe novel MS2-aligning spacers sampled from throughout the MS2 genome (Fig. 3c–e, Extended Data Fig. 5c–f). The MS2-aligning spacers shared no sequence similarity with the plasmid or host genome, confirming their specificity (Extended Data Fig. 5d). In summary, *Fs*RT–Cas1–Cas2 enables spacer acquisition directly from a foreign RNA, thereby providing a molecular memory of an invading virus.

Recording of arbitrary transcripts using Record–seq

To assess the potential of *Fs*RT–Cas1–Cas2 for quantitatively recording transcriptional events, we used an inducible expression system to directly determine whether spacers were being acquired according to RNA abundance. The corresponding constructs contained *super-folder GFP* (*sfGFP*) or *renilla luciferase* (*Rluc*) genes under the transcriptional control of the anhydrotetracycline (aTc)-inducible P_{tetA} promoter. We introduced these constructs into *E. coli* cultured in increasing levels of aTc and subsequently collected both total RNA and plasmid DNA for quantitative PCR with reverse transcription and Record–seq, respectively (Fig. 3f). Upon increasing induction of sf*GFP* or *Rluc*, there was a concordant dose-dependent increase in the coverage of

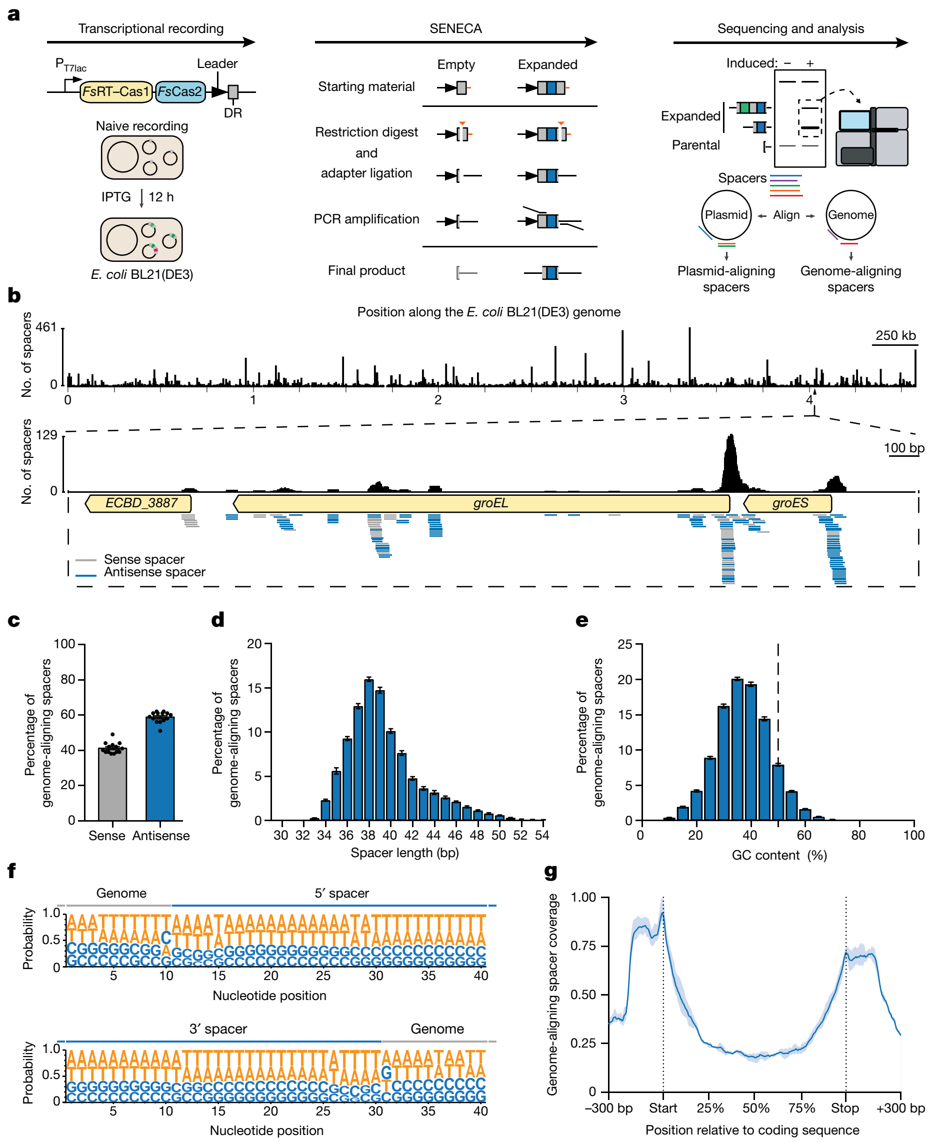


Fig. 2 | Characterization of spacers acquired by *Fs*RT–Cas1–Cas2. **a**, Schematic of Record–seq experimental workflow (Extended Data Fig. 2a). DR, direct repeat. **b**, Coverage of spacers aligning to the *E. coli* genome and a representative locus. Identical alignments represent recurrent spacers acquired in independent biological samples ($n = 14$). **c**, Spacers aligning to sense or antisense strand of coding genes in the *E. coli* genome. The sense/antisense orientation label in **c**, **d** is with respect to the RNA. **d**, Length distribution of genome-aligning spacers. **e**, GC content distribution of genome-aligning spacers. Dotted line represents 50% GC content. **f**, Nucleotide probabilities of the 5′ (left) or 3′ (right) end of the spacer, along with the respective flanking sequence. Spacer (blue) and flanking (grey) nucleotides are shown. Data represent spacers merged across $n = 14$ independent biological samples. **g**, Gene body coverage of spacer alignments along transcripts. Relative position represents percentiles of coding sequence lengths ± 300 bp of adjacent genomic regions. Values are mean normalized coverage ± s.d., $n = 14$ independent biological samples. Values in **c**–**e** are mean percentage of genome-aligning spacers (with individual data points for **c**, ± s.e.m. for **d**, **e**), $n = 14$ independent biological samples.

spacers aligning to the respective coding sequence (Extended Data Fig. 6a). We quantified this response and observed a linear relationship ($R^2 = 0.97$) between spacer counts and absolute mRNA copy number (Extended Data Fig. 6b–e) or aTc concentration in the medium (Fig. 3f). Furthermore, *sfGFP*-aligning spacers were readily detected against the backdrop of genome-aligning spacers by almost an order of magnitude (Extended Data Fig. 6f, g), which is in line with using a strong synthetic inducible promoter such at P_{tetA}. Notably, spacers that aligned to the constitutively expressed *KanR* gene were not dependent on aTc concentration (Fig. 3f).

To generalize these findings further, we evaluated a second inducible expression system, placing the *firefly luciferase* (*Fluc*) gene downstream of the 3-oxohexanoyl-homoserine lactone (3OC6-HSL)-inducible P_{LuxR} promoter. Induction led to a fourfold increase in *Fluc*-aligning spacers (Extended Data Fig. 6h). Furthermore, combining both the aTc-inducible P_{tetA} and the 3OC6-HSL-inducible P_{LuxR} transcription system enabled orthogonal recording of two independent stimuli in parallel (Fig. 3g, Extended Data Fig. 6i, j). This suggests that Record–seq is compatible with seemingly any inducible expression system, thereby enabling recording of multiple orthogonal sets of defined stimuli within a population of living cells. Together, these results show that CRISPR spacer acquisition from RNA can generate a quantifiable record of cumulative transcript abundance, and also that the transcriptional records are efficiently retrieved using standard molecular and sequencing methods.

Record–seq shows cumulatively highly expressed genes

Considering that *Fs*RT–Cas1–Cas2 acquired spacers directly from RNA in an abundance-dependent manner, we investigated whether this could enable quantification of the cumulative cellular transcriptome.

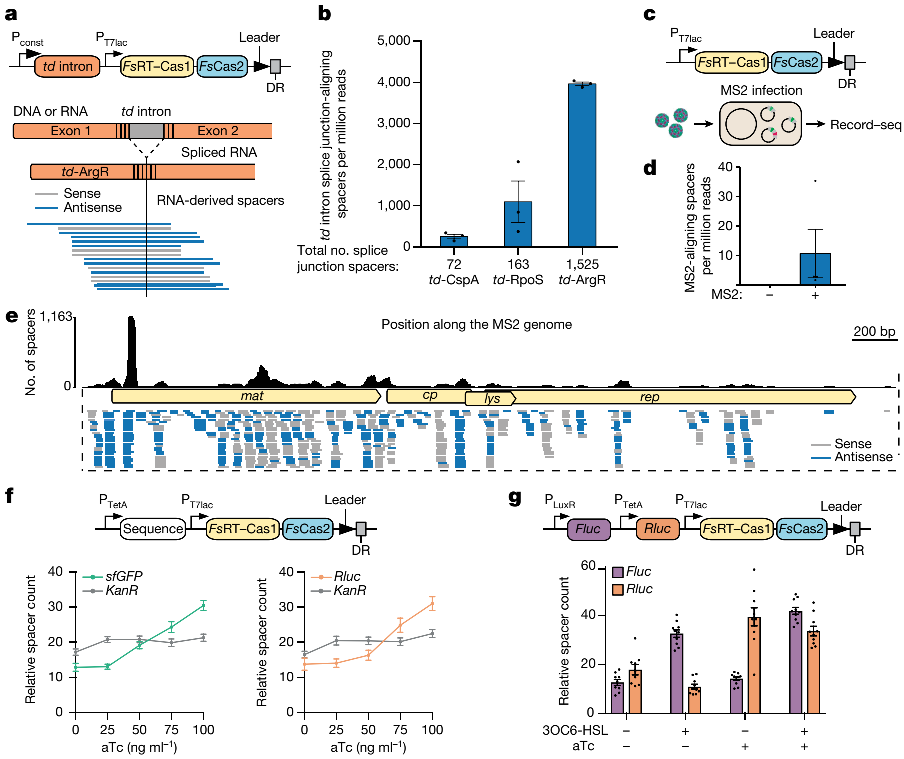


Fig. 3 | *Fs*RT–Cas1–Cas2 acquires spacers directly from RNA according to abundance. **a**, Schematic of *td* intron-containing constructs and representative spacers aligning to the *td* intron splice junction. **b**, Quantification of spacers derived from the *td* intron splice junction. Values are mean ± s.e.m. *td* intron spacers per million reads, $n = 3$ independent biological samples. The sum of raw sequencing counts is shown below. **c**, Experimental workflow depicting MS2 recording. **d**, Quantification of MS2-derived RNA spacers. Values are mean ± s.e.m. MS2-aligning spacers per million reads, $n = 3$ (no MS2) and 4 (MS2) biologically independent samples. **e**, Coverage of spacers aligning to the MS2 genome. Data represent alignments merged across samples. Sense or antisense orientation is given with respect to the plus-strand MS2 RNA. **f**, Schematic and quantification of transcriptional recording of arbitrary sequences. Values are mean ± s.e.m. relative spacer count, $n = 10$ independent biological samples. The constitutively expressed *KanR* selection marker was used as a control. **g**, Schematic and quantification of orthogonal transcriptional recording. Values are mean ± s.e.m. relative spacer count, $n = 10$ (treated) and 9 (untreated) independent biological samples.

We collected both plasmid DNA for Record–seq and total RNA for RNA-seq from cultures of *E. coli* overexpressing *Fs*RT–Cas1–Cas2 (Fig. 4a). First, we confirmed the reproducibility of Record–seq between biological replicates (Pearson correlation, 0.996–0.999; $R^2 = 0.560$–0.618) (Extended Data Fig. 7a), and then we assessed the influence of gene expression on spacer acquisition. The *Fs*RT–Cas1–Cas2 spacers showed a strong bias towards highly transcribed genes (Extended Data Fig. 7a) and correlated with RNA-seq-based gene expression values transcriptome-wide at various growth stages (Extended Data Fig. 7b–d). Although certain CRISPR–Cas subtypes possess active mechanisms for preferentially acquiring plasmid-derived spacers[33], we did not observe the same after accounting for the high expression level of these genes (Extended Data Fig. 7e). Thus, spacers are systematically acquired from highly transcribed genes, and represent cumulative transcript expression.

Transcriptome-scale recording reveals cell behaviours

To determine whether Record–seq could be used to record and describe complex cellular behaviours, we turned to the well-studied oxidative stress and acid stress responses in *E. coli*. We performed Record–seq on oxidative and acid stress-stimulated[34,35] cultures expressing *Fs*RT–Cas1–Cas2 and analysed cumulative expression counts using unsupervised hierarchical clustering and principal component analysis (PCA). Both approaches were successful in distinguishing treatment conditions, suggesting that Record–seq captured the differential molecular histories (Fig. 4b–e). To identify the cumulatively differentially expressed genes, we leveraged standard differential expression analysis tools developed for RNA sequencing. To overcome specific biases and assumptions of individual tools, we used three complementary tools, namely DESeq2[36], edgeR[37], and baySeq[38]. After identifying differentially expressed genes with each tool (Supplementary Table 9), we generated a set of signature genes for each stimulus based on the union of the top 20 differentially expressed genes from each analysis, which we hierarchically clustered and plotted along with their expression values (Fig. 4f, g). Among the signature genes, we identified several that were expected to dominate the cellular responses for each stimulus (Supplementary Notes). We investigated the minimum number of cells required for assessing complex cellular behaviours by Record–seq, finding that 8.8×10^6 cells are sufficient to appropriately classify treatment conditions (Extended Data Figs. 8, 9, Supplementary Notes). In summary, these data support the notion that the RNA-derived spacers stored within CRISPR arrays can be used to reconstruct the transcriptional response underlying a complex cellular behaviour.

Sentinel cells encode transient herbicide exposure

To determine whether Record–seq could be leveraged to produce sentinel cells, we used the herbicide paraquat and tested whether Record–seq could capture dose-dependent and transient exposure. Paraquat is a bacteriostatic herbicide that results in superoxide anion production in microbes[39,40], and is banned in a number of countries owing to its acute toxicity in humans and use in suicide[41].

Using an improved *Fs*RT–Cas1–Cas2 expression construct (Extended Data Fig. 10a, b), we exposed *E. coli* cultures to increasing concentrations of paraquat and retrieved the transcriptional memories by Record–seq. Quantification of cumulative gene expression under the different treatment conditions showed that samples were readily classified into appropriate exposure groups using both unsupervised hierarchical clustering and PCA (Fig. 5a, b). Moreover, the signature genes captured dose-responsive and canonical paraquat-exposure genes within *E. coli* (Fig. 5c). For example, within the signature genes we found *ahpC* and *ahpF*, which encode the two subunits of an alkyl hydroperoxide reductase previously shown to facilitate scavenging of reactive oxygen species (ROS) caused by paraquat[42]. In addition, we identified a set of genes of the cys-regulon involved in cysteine metabolism, namely *cysC*, *cysJ* and *cysK*, which have been shown to facilitate paraquat resistance in *E. coli*[42,43].

We next determined whether Record–seq could also capture transient paraquat exposure in a physiological range[41]. After transiently stimulating cultures with paraquat (Fig. 5d), we quantified cumulative gene expression and gene expression for Record–seq and RNA-seq data sets, respectively. Then, we assessed whether the two methods could capture the transient paraquat exposure using PCA (Fig. 5e, f) and differentially expressed signature gene clustering (Extended Data Fig. 10c, d). These analyses show that Record–seq, but not RNA-seq, was capable of capturing the transient paraquat exposure (Fig. 5e, f and Extended Data Fig. 10c, d). Together, these results show that the

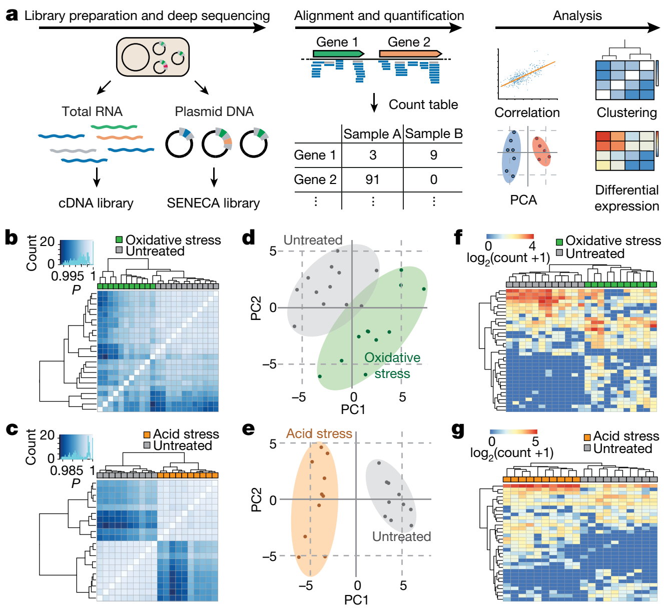


Fig. 4 | Transcriptome-scale recording and analysis of complex cellular behaviours. a, Workflow for comparing Record–seq with RNA-seq. See Supplementary Notes for further discussion. **b**, Clustering of Record–seq data from untreated (grey) and oxidative-stress-treated (green) *E. coli* populations, performed using Pearson correlation, $n = 12$ (untreated) and $n = 11$ (treated) independent biological samples. **c**, Clustering of Record–seq data from untreated (grey boxes) and acid-stress-treated (orange boxes) *E. coli* populations, performed using Pearson correlation, $n = 10$ independent biological samples. **d**, PCA of Record–seq data from untreated (grey) and oxidative-stress-treated (green) *E. coli* populations, $n = 12$ (untreated) and $n = 11$ (treated) independent biological samples. **e**, PCA of Record–seq data from untreated (grey) and acid-stress-treated (orange) *E. coli* populations, $n = 10$ independent biological samples. **f**, Clustering of Record–seq data for signature differentially expressed genes under oxidative stress. **g**, Clustering of Record–seq data for signature differentially expressed genes under acid stress.

memory of paraquat exposure was lost within the cellular transcriptome as assessed by RNA-seq, but preserved within the molecular memories stored within the DNA of the CRISPR arrays of the sentinel cells, as investigated by Record–seq.

Discussion

Here, we have described Record–seq, a technique for encoding transcriptome-scale events into DNA and assessing the cumulative gene expression of populations of cells. We have demonstrated its potential by recording specific and complex transcriptional information. First, to improve on existing spacer readout methods, we developed SENECA, resulting in a several thousand-fold improvement in spacer detection efficiency compared to recent reports[26,44,45], thereby enabling in-depth characterization of *Fs*RT–Cas1–Cas2 and its application as a molecular recorder. Our results suggest that RNA-derived spacers are preferentially acquired from the ends of abundant transcripts from AT-rich regions with no PAM, and are broadly sampled at the transcriptome-scale, enabling the parallelized quantification of cumulative transcript expression.

In a set of experiments, we have shown that upon increasing induction of arbitrary sequences, spacers are acquired in an orthogonal, dose-dependent manner and highly correlate with the absolute mRNA copy number in the cell, thus demonstrating that the molecular record faithfully recapitulates the initial stimulus in a predictable way. This also paves the way for increasingly multiplexed and orthogonal molecular recording devices. Upon inducing complex cellular behaviours, Record–seq provides a meaningful transcriptome-scale record of molecular events, which exceeds the capabilities of current molecular recording technologies that only record specific stimuli[15–25]. Finally, we used Record–seq to elucidate dose-dependent features of the complex cellular response to the bacteriostatic herbicide paraquat, and have shown that Record–seq, but not RNA-seq, can record transient paraquat stimulation.

Although additional work will greatly improve the capacity of Record–seq to encode richer and more dynamic expression and lineage information within fewer cells (Supplementary Notes), our proof-of-principle experiments introduce a powerful tool for recording transcriptome-scale events permanently in DNA for later

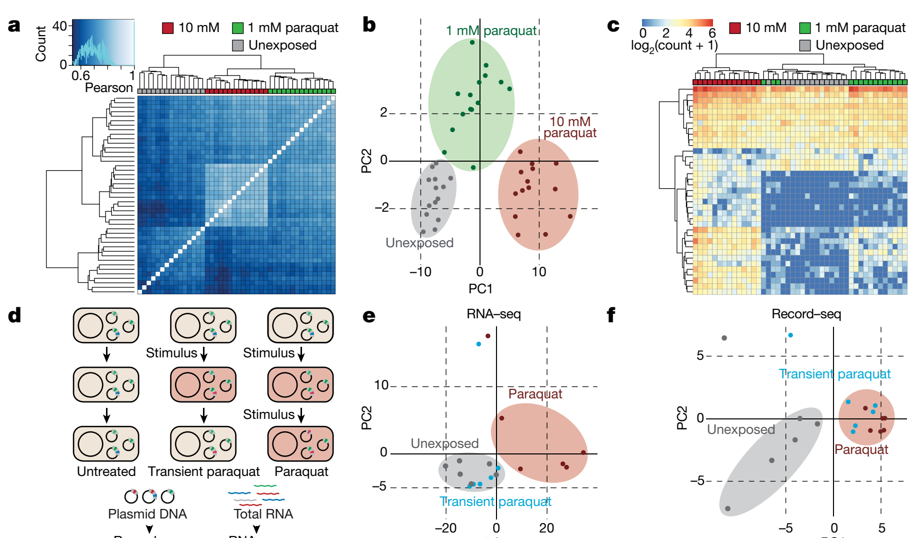


Fig. 5 | Sentinel cells for recording of dose-dependent and transient herbicide exposure. a, Clustering of Record–seq data from untreated (grey), 10 mM paraquat-treated (red) and 1 mM paraquat-treated (green) *E. coli* populations, performed using Pearson correlation, $n = 15$ independent biological samples. **b**, PCA of Record–seq data from untreated (grey), 10 mM paraquat-treated (red) and 1 mM paraquat-treated (green) *E. coli* populations, $n = 15$ independent biological samples. **c**, Clustering of Record–seq data for signature differentially expressed genes. **d**, Workflow for comparing Record–seq with RNA-seq for analysis of transient paraquat exposure. **e**, PCA of RNA-seq data from unexposed (grey), transiently paraquat-exposed (turquoise) and constantly paraquat-exposed (red) *E. coli* populations, $n = 6$ independent biological samples. **f**, PCA of Record–seq data from unexposed (grey), transiently paraquat-exposed (turquoise) and constantly paraquat-exposed (red) *E. coli* populations, $n = 6$ independent biological samples.

reconstruction of complex molecular histories from populations of cells. The recorded transcriptional histories reflect the underlying gene expression changes and could therefore be used to interrogate biological or disease processes. In the long term, we envision that CRISPR spacer acquisition components could be introduced into other cell types to record the molecular sequences of events, and lineage paths, that gives rise to particular cell behaviours, cell states and types.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0569-1.

Received: 2 May 2018; Accepted: 21 August 2018;
Published online 3 October 2018.

1. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nat. Rev. Genet.* **19**, 299–310 (2018).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
3. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).
4. Schmidt, F. & Platt, R. J. Applications of CRISPR–Cas for synthetic biology and genetic recording. *Curr. Opin. Syst. Biol.* **5**, 9–15 (2017).
5. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
6. Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182 (2005).
7. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
8. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663 (2005).
9. Garneau, J. E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
10. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **12**, 479–492 (2014).
11. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181–190 (2010).
12. Amitai, G. & Sorek, R. CRISPR–Cas adaptation: insights into the mechanism of action. *Nat. Rev. Microbiol.* **14**, 67–76 (2016).
13. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
14. Sternberg, S. H., Richter, H., Charpentier, E. & Qimron, U. Adaptation in CRISPR–Cas systems. *Mol. Cell* **61**, 797–808 (2016).
15. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016).
16. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345–349 (2017).
17. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Science* **358**, 1457–1461 (2017).
18. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR–Cas in human cells. *Science* **353**, aag0511 (2016).
19. Frieda, K. L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
20. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* **360**, eaap8992 (2018).
21. Farzadfard, F. & Lu, T. K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346**, 1256272 (2014).
22. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
23. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
24. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat. Methods* **14**, 195–200 (2017).
25. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
26. Silas, S. et al. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase–Cas1 fusion protein. *Science* **351**, aad4234 (2016).
27. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576 (2012).
28. Erdmann, S., Le Moine Bauer, S. & Garrett, R. A. Inter-viral conflicts that exploit host CRISPR immune systems of *Sulfolobus*. *Mol. Microbiol.* **91**, 900–917 (2014).
29. Pyenson, N. C., Gayvert, K., Varble, A., Elemento, O. & Marraffini, L. A. Broad targeting specificity during bacterial type III CRISPR–Cas immunity constrains viral escape. *Cell Host Microbe* **22**, 343–353 (2017).
30. Sandegren, L. & Sjöberg, B.-M. Self-splicing of the bacteriophage T4 group I introns requires efficient translation of the pre-mRNA in vivo and correlates with the growth state of the infected bacterium. *J. Bacteriol.* **189**, 980–990 (2007).
31. Belfort, M. et al. Processing of the intron-containing thymidylate synthase (*td*) gene of phage T4 is at the RNA level. *Cell* **41**, 375–382 (1985).
32. Gott, J. M., Shub, D. A. & Belfort, M. Multiple self-splicing introns in bacteriophage T4: evidence from autocatalytic GTP labeling of RNA in vitro. *Cell* **47**, 81–87 (1986).
33. Levy, A. et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510 (2015).
34. Zheng, M. et al. DNA microarray-mediated transcriptional profiling of the *Escherichia coli* response to hydrogen peroxide. *J. Bacteriol.* **183**, 4562–4570 (2001).
35. Maurer, L. M., Yohannes, E., Bondurant, S. S., Radmacher, M. & Slonczewski, J. L. pH regulates genes for flagellar motility, catabolism, and oxidative stress in *Escherichia coli* K-12. *J. Bacteriol.* **187**, 304–319 (2005).
36. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
37. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
38. Hardcastle, T. J. & Kelly, K. A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 422 (2010).
39. Hassan, H. M. & Fridovich, I. Paraquat and *Escherichia coli*. Mechanism of production of extracellular superoxide radical. *J. Biol. Chem.* **254**, 10846–10852 (1979).
40. Ochsner, U. A., Vasil, M. L., Alsabbagh, E., Parvatiyar, K. & Hassett, D. J. Role of the *Pseudomonas aeruginosa oxyR–recG* operon in oxidative stress defense and DNA repair: OxyR-dependent regulation of *katB–ankB*, *ahpB*, and *ahpC–ahpF*. *J. Bacteriol.* **182**, 4533–4544 (2000).
41. Wesseling, C., Corriols, M. & Bravo, V. Acute pesticide poisoning and pesticide registration in Central America. *Toxicol. Appl. Pharmacol.* **207** (Suppl.), 697–705 (2005).
42. Pomposiello, P. J., Bennik, M. H. & Demple, B. Genome-wide transcriptional profiling of the *Escherichia coli* responses to superoxide stress and sodium salicylate. *J. Bacteriol.* **183**, 3890–3902 (2001).
43. Fuentes, D. E. et al. Cysteine metabolism-related genes and bacterial resistance to potassium tellurite. *J. Bacteriol.* **189**, 8953–8960 (2007).
44. Silas, S. et al. Type III CRISPR–Cas systems can provide redundancy to counteract viral escape from type I systems. *eLife* **6**, e27601 (2017).
45. Silas, S. et al. On the origin of reverse transcriptase—using CRISPR–Cas systems and their hyperdiverse, enigmatic spacer repertoires. *MBio* **8**, e00897-17 (2017).

Acknowledgements We thank M. Okoniewski for assistance with data analysis; S. Ghosh and T. Tanna for technical assistance; S. Silas, A. Z. Fire, and the entire Platt Laboratory for discussions; S. Panke, M. Jeschek, L. Pestalozzi, I. Wüthrich, and D. Gerngross for reagents and comments; C. Beisel, E. Burcklen, K. Eschbach, I. Nissen, and M. Kohler from the Genomics Facility Basel for assistance in Illumina sequencing. R.J.P., M.Y.C and F.S. are supported, in part, by funds from the Swiss National Science Foundation, ETH domain Personalized Health and Related Technologies, Brain and Behavior Research Foundation, and the National Centres of Competence – Molecular Systems Engineering.

Reviewer information *Nature* thanks C. Beisel and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions F.S. and M.Y.C. performed the experiments; F.S., M.Y.C. and R.J.P. analysed the data; and F.S., M.Y.C. and R.J.P. wrote the manuscript.

Competing interests Patent applications have been filed relating to work in this manuscript.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0569-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0569-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to R.J.P.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Orthologue discovery pipeline. The protein sequence of *Arthrospira platensis* RT–Cas1 (WP_006620498) was used as a seed sequence, and a JACKHMMER search was run against all NCBI non-redundant protein sequences using HMMER v3.1b2 (*E*-value cutoff of 1×10^{-5})[46]. Proteins with both Cas1 and reverse transcriptase domains were subsequently identified using HMMSCAN (*E*-value cutoff of 1×10^{-5}). Genome sequence information for the candidate proteins was retrieved and further inspected for the presence of RT–Cas1, Cas2, and a CRISPR array using CRISPRdetect v2.0[47], CRISPRone[48], and HMMSCAN[46]. From 121 candidate proteins, 14 CRISPR loci were selected and subsequently aligned using MUSCLE v3.8.31[49] to identify candidate domains and catalytic residues. Genetic distances were computed using the Jukes–Cantor method and a phylogenetic tree was built using the nearest-neighbour method.

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Bacterial strains and culture conditions. *E. coli* strains used in this study were Stbl3 (Thermo Fisher Scientific) for cloning purposes, BL21(DE3) Gold (Agilent Technologies), BL21AI (Invitrogen) and NovaBlue(DE3) (EMD Millipore) as a K12 strain for acquisition assays. All strains were made competent using the Mix & Go *E. coli* Transformation Kit & Buffer Set (Zymo Research) following the manufacturer's protocol with growth in ZymoBroth at 19 °C directly from fresh colonies. After transformation, cells were grown at 37 °C on lysogenic broth (LB) (Difco) 1.5% agar plates containing 50 μg/ml kanamycin and 1% glucose (w/v) to reduce background expression from the T7lac system. Liquid cultures for plasmid isolation were grown in TB medium (24 g/l yeast extract, 20 g/l tryptone, 4 ml/l glycerol, 17 mM KH_2PO_4, 72 mM K_2HPO_4) containing 1% glucose (w/v).

Generation of Golden Gate-compatible pET30 overexpression vector. All standard PCRs for cloning were performed using Phusion Flash High-Fidelity PCR Master Mix (Thermo Scientific) or KAPA HiFi HotStart ReadyMix (Roche), oligonucleotides and gBlocks were ordered from Integrated DNA technologies. Primers are listed in Supplementary Table 7. pET30b(+) (kind gift from M. Jeschek) was PCR amplified as five fragments using primers FS_151/FS_152, FS_153/FS_154, FS_155/FS_156, FS_157/FS_158, FS_159/FS_160, respectively, to remove the five undesired BbsI restriction sites present in the backbone. The resulting PCR fragments were assembled using 2 × HiFi DNA Assembly Mastermix (NEB), yielding pFS_0012. Subsequently, oligos FS_380 and FS_381 were annealed to generate a double-stranded DNA (dsDNA) fragment encoding the T7 terminator and cloned into pFS_0012 using XhoI/CsiI, yielding pFS_0013—a pET30-derived overexpression vector harbouring two Golden Gate cloning sites and thus facilitating parallel cloning of RT–Cas1, Cas2 and a corresponding CRISPR array. Nucleotide sequences of all RT–Cas1 and Cas2 orthologues tested in this study along with their corresponding CRISPR arrays are listed in the Supplementary Information, Sequences.

Golden Gate assembly of RT–Cas1–Cas2 overexpression vectors for orthologue screen. RT–Cas1, Cas2 and CRISPR array sequences were ordered from Twist Biosciences and Genscript. Putative CRISPR arrays were ordered as sequences consisting of the leader sequence followed by DR–nativespacer1–DR–nativespacer2–DR. Furthermore, each fragment was flanked by BbsI restriction sites generating overhangs facilitating Golden Gate Assembly into pFS_0013. In brief, 40 fmol per fragment (RT–Cas1, Cas2, corresponding CRISPR array, pFS_0013 acceptor vector), 1 μl ATP/DTT mix (10 mM each), 0.25 μl T7 DNA Ligase (Enzymatics), 0.75 μl BpiI (Thermo Scientific), 1 μl buffer green up to 10 μl with PCR grade H_2O were subjected to 99 cycles of 37 °C for 3 min, 16 °C for 5 min, followed by 80 °C for 10 min. Subsequently, 5 μl of this mixture was transformed into 50 μl Stbl3 cells and recovered in SOC medium for 30 min at 37 °C, 1,000 r.p.m. before spreading on plates.

Spacer acquisition. Acquisition assays were performed at 37 °C, 300 r.p.m. in bacterial culture tubes containing 3 ml TB medium supplied with 100 μM isopropyl-β-D-thiogalactopyranoside (IPTG) (Sigma Aldrich) for BL21(DE3) Gold and NovaBlue(DE3). For *E. coli* BL21AI, L-(+)-arabinose (Sigma Aldrich) was additionally added to 0.2% (w/v). Each culture was inoculated with two colonies of bacteria stored no longer than 14 days at 4 °C upon transformation and overnight growth at 37 °C. When cultures reached saturation (typically 12–14 h after inoculation), 2 ml of bacterial culture was harvested and plasmids containing CRISPR arrays were isolated by standard plasmid Mini-Prep procedures to serve as a template for preparation of deep sequencing libraries.

Amplification of CRISPR arrays for classical acquisition readout by deep sequencing. Leader proximal spacers were PCR amplified from 3 ng plasmid DNA per μl PCR reaction using NEBNext High-Fidelity 2× PCR Master Mix (NEB) with a forward primer binding in the leader sequence of the respective CRISPR array and a reverse primer binding in the first native spacer (see Supplementary Information (Primer Design Note 1) and Supplementary Table 3 for primer design and binding sites of individual CRISPR arrays, respectively). For each biological replicate, 12 individual PCR reactions of 10 μl were performed with an extension time of 15 s for 16 cycles. The individual 10-μl reactions belonging to the same biological sample were then pooled, and residual primers removed using in-house-generated AMPure beads at a PCR to bead ratio of 1:1.5 (v/v) eluting the PCR product in 60 μl of buffer TE. Subsequently, 500 ng of first-round PCR product per biological sample was run on a 3% laboratory agarose gel (300 V, 55 min, cooling the gel-chamber in an ice-water bath during the run) and purified by blind excision of gel slices between 211 and 300 bp, avoiding the prominent DNA band corresponding to PCR products of the unexpanded array (that is, no acquisition of novel spacers). Amplicons were then purified from the gel slices using the QIAquick Gel Extraction Kit (QIAGEN) and eluted into 22 μl buffer EB. Illumina sequencing adaptors and indices were appended in a second round of PCR, using 6 μl gel-purified input DNA as a template in a 20-μl PCR reaction with universal second-round deep sequencing primers attaching P5 and P7 handles for binding of PCR products to the flow cell in deep sequencing as well as barcoding the samples with $(N)_8$ barcodes corresponding to Illumina TruSeq HT indices (Supplementary Information (Primer Design Note 2) and Supplementary Table 4 for primer design and indices, respectively). After this second round of PCR, products were purified using the QIAquick PCR Purification Kit (QIAGEN) and eluted in 22 μl buffer EB. Samples were then pooled and subjected to another round of gel purification using the same parameters as described above, this time excising products in the range of 280–350 bp.

Selective amplification of ExpaNdEd CRISPR arrays (SENECA). *Fs*CRISPRArray2 was amplified from pFS_160 using FS_871/FS_904, generating a minimal *Fs* CRISPR array consisting of the leader sequence and a single direct repeat followed by a FaqI restriction site (CTTCAG) on the bottom strand resulting in plasmid pFS_0235 as our standard recording plasmid. This plasmid was transformed into chemocompetent BL21(DE3) Gold bacteria or NovaBlue(DE3) (EMD Millipore) and subjected to spacer acquisition as described above. Following plasmid extraction and quantification using Quant-IT PicoGreen dsDNA Assay Kit (Thermo Scientific) read out with a Tecan M1000 Pro Microplate reader, plasmid DNA was subjected to SENECA-adaptor ligation in a Golden Gate reaction. Oligonucleotides FS_0963/FS_0964 were annealed (2.5 μl each of 100 μM oligo, 5 μl NEBuffer 2 (NEB), 40 μl PCR grade H_2O), by heating to 95 °C for 5 min and cooling to 20 °C at 0.12 °C per s. Annealed oligos were diluted 1:100 in TE buffer. Next, 40 fmol plasmid DNA (180.3 ng for pFS_0235), 0.25 μl T7 ligase (Enzymatics), 1 μl FastDigest FaqI, 0.5 μl of 20 × SAM, 1 mM ATP, 1 mM DTT (all Thermo Scientific), 1 μl of annealed, diluted oligonucleotides FS_0963/FS_0964 in 10 μl total volume were subjected to 99 cycles of 3 min 37 °C, 3 min 20 °C followed by 15 min at 55 °C. First-round deep sequencing PCR was performed using NEBNext High-Fidelity 2× PCR Master Mix (NEB) (forward primers: FS_0968 to FS_0974, reverse primer: FS_0911). For each biosample, one 30-μl reaction containing 10.38 μl adaptor ligated plasmid DNA was performed (98 °C for 30 s; 22 cycles at 98 °C for 10 s, 57 °C for 30 s and 72 °C for 20 s followed by 72 °C for 5 min), pooled and purified using magnetic beads (GE Healthcare) at a PCR-to-bead ratio of 1:1.6 (v/v) recovering the PCR product in 25 μl TE buffer (Supplementary Information (Primer Design Note 3) for details on primer design). Illumina sequencing adaptors and indices were appended in a second round of PCR (98 °C for 30 s, 8 cycles of 98 °C for 10 s, 65 °C for 30 s and 72 °C for 30 s, and 72 °C for 5 min) using 5 μl first-round PCR product as input in a 20-μl reaction (Supplementary Information (Primer Design Note 2) and Supplementary Table 4 for primer design and indices, respectively). Samples were pooled, desalted using the QIAquick PCR Purification Kit (QIAGEN) and size selected on a E-Gel EX Agarose Gels, 2% (Thermo Scientific), loading 200–500 ng DNA per lane, extracted using the QIAquick Gel Extraction Kit and subjected to deep sequencing on Illumina MiSeq or NextSeq500 platforms using the MiSeq Reagent Kit v3 (150 cycles) or NextSeq 500/550 Mid/High Output v2 kit (150 cycles) (both Ilumina), respectively. Libraries were loaded at a concentration of 1.4 to 1.6 pM as determined by quantitative PCR with reverse transcription (qRT–PCR) using the KAPA Library Quantification Kit for Illumina Platforms (Roche). PhiX was included at 5–10%.

SENECA-based orthologue screen. For the SENECA-based CRISPR array directionality screen, putative CRISPR arrays were extracted from genomic sequences, assuming a standard leader length of 150 nucleotides (nt) followed by a single direct repeat. The FaqI restriction site required for SENECA was appended downstream of the direct repeat and sequences were flanked by universal adapters for amplification and cloning. The final array sequences including these features are provided in the Supplementary Information (Supplementary Information, Sequences 2) and were ordered from Twist Biosciences as linear DNA fragments. These were PCR amplified using primers FS_1406/FS_1407 and cloned into CsiI/NotI-digested plasmids containing their respective RT–Cas1–Cas2 orthologue using HiFi DNA Assembly (NEB). Upon transformation into *E. coli* BL21(DE3), these constructs were subjected to the standard spacer acquisition assay in TB medium. Plasmid DNA was extracted and subjected to SENECA adaptor ligation. The respective

oligos to be annealed for each CRISPR array tested in this experiment are listed in Supplementary Table 5. Following adaptor ligation, a single 140 μl first-round PCR reaction was prepared for each orthologue using NEBNext High-Fidelity 2× PCR Master Mix and containing the entire 20-μl SENECA adaptor ligation as a template. First-round PCR primers specific to the respective direct repeat of each CRISPR array tested are listed in Supplementary Table 6. The 140-μl PCR reaction was split into 12 reactions of 11 μl along the row of a 96-well plate. This plate was subjected to a gradient PCR (53 to 68 °C in an Eppendorf Mastercycler Gradient). This procedure was chosen, because SENECA leverages the fact that a direct repeat matching primer will only bind to the full direct repeat resulting from an acquisition event but not the truncated parental direct repeat at a unique annealing temperature. By splitting the PCR reaction and subjecting it to a temperature gradient, it is ensured that without a prior knowledge, at least one of the 12 reactions is subjected to the annealing temperature at which selective amplification of expanded CRISPR arrays occurs. PCR was performed for 30 cycles upon which, the 12 reactions performed along the temperature gradient were pooled again and purified using 1.85× Ampure beads and eluted in 25 μl TE buffer. Then, 5 μl of this elution was used as a template for a standard 20-μl second-round PCR at 65 °C annealing temperature for 12 cycles as described above. Subsequently, PCR products were purified using 2.2× Ampure beads, eluted into 22 μl TE buffer, size selected as described in the standard SENECA protocol (E-Gel Ex 2%, followed by gel extraction) and subjected to deep sequencing.

Deep sequencing. Small-scale targeted deep sequencing of CRISPR arrays for the orthologue screen was performed using the Illumina MiSeq v3 300 cycle kit on an Illumina MiSeq platform or Illumina HiSeq High Output High Output PE 200 cycle kit an Illumina HighSeq2500. Deep sequencing of spacer libraries prepared using SENECA were sequenced using the NextSeq 550/550 High Output Kit v2 150 cycle on Illumina NextSeq platform or the MiSeq Reagent Kit v3 150-cycle on a MiSeq.

Data analysis pipeline. FASTQ files were quality filtered and trimmed using trimmomatic (trimmomatic s.e. LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:75)[50] and subsequently converted to FASTA files using FASTX-Toolkit v0.0.14 (fastq-to-fasta) (http://hannonlab.cshl.edu/fastx_toolkit/). Using custom scripts written in Python v2.7, spacers were identified based on the identification of a 20–66 nucleotide sequence between two 10-nt direct repeat segments, allowing for 2 and 3 mismatches in the first and second direct repeat segments, respectively. Arrays with multiple spacers were identified based on the presence of a complete direct repeat sequence, allowing for 3 mismatches. Only unique spacers (>1 mismatch) from a given sample were processed further. Spacers were aligned to a merged reference genome containing plasmid and *E. coli* sequences (*E. coli* Bl21(DE3) Gold (NC_012947.1) genome, *E. coli* K12 (NC_000913.3)) using Bowtie2 (bowtie2–very-sensitive-local)[51]. In MS2 challenge experiments, the MS2 sequence (MS2 (NC_001417.2)) was also included in the merged reference genome. Identical alignments were collapsed using Samtools v1.3[52], and alignments were visualized in Geneious v10.2.3. Basic statistics about numbers of reads or alignment features were calculated using standard bash commands, and compiled and visualized using Prism v7.0d. Gene body percentiles were calculated using RSeQC (geneBody_coverage.py v2.6.4)[53]. Nucleotide probabilities were determined and visualized using the weblogo webtool v2.8.2[54]. Simulated spacer data sets were prepared using BEDtools v2.25 (bedtools random -n 500 -l 38)[55]. Transcript quantification for RNA-seq and Record–seq was performed using featureCounts v1.5.0[56]. Using custom scripts written in MATLAB v9.1.0, RNA-seq and Record–seq transcript counts were normalized using transcripts per million (TPM) and used to compute cumulative spacer sums, a linear regression fit, coefficient of determination (R^2), and Pearson linear correlation coefficient.

Record–seq data sets corresponding to oxidative or acid stress treatments were analysed using custom scripts written in R v3.4.4. In brief, transcripts with fewer than 5 counts across replicates were discarded. Heatmaps representing unsupervised hierarchical clustering of Pearson linear correlation with complete linkage (using raw transcript counts as inputs) were prepared using the 'heatmap.2', 'hclust', and 'cor' commands with default settings. PCA was performed on $\log_2$ transformed data (raw counts plus one pseudocount to tolerate zeros) for the 50 most variable (s.d.) genes using the 'prcomp' command with default settings. Differential expression analyses (using raw counts plus one pseudocount as input) were performed using DEseq2 v1.14.1, edgeR v3.16.5, and baySeq v2.8.0 encapsulations within R. Heatmaps representing unsupervised hierarchical clustering of signature differentially expressed genes were prepared using the 'pheatmap' command with default settings.

RNASeq of *E. coli* BL21(DE3). RNA extraction from *E. coli* BL21(DE3) was performed after overnight growth under induction of *Fs*RT–Cas1–Cas2 expression following the QIAGEN Supplementary Protocol: Purification of total RNA from bacteria using the RNeasy Mini Kit. To achieve the appropriate amount of input culture (corresponding to 5×10^8 cells), serial dilutions of the overnight culture were prepared to achieve an OD_{600} between 0.2 and 0.6 measured with a NanoDrop OneC (Thermo Scientific). Bacteria were lysed using acid-washed glass beads (G1277-10G, Sigma Aldrich). The additional on-column DNase digestion was performed using the RNase-Free DNase Set (QIAGEN). DNA-free RNA was submitted to the Genomics Facility Basel for ribosomal RNA (rRNA) depletion using the Ribo-Zero rRNA Removal Kit (Illumina) and followed by library preparation and sequencing on an Illumina NextSeq platform using the NextSeq 500/550 High Output v2 kit (150 cycles).

***td* intron.** The gBlock FS_gBlock_*td*_intron_acceptor (Supplementary Information, Sequences 3) was cloned into pFS_0235 using SphI/SgrAI yielding pFS_0238. This gBlock encoded the BBa_J23104 promoter, the ribosome binding site from bacteriophage T7 gene 10 as well as the *td* intron sequence including flanking regions facilitating efficient splicing. Furthermore, a BbsI-mediated Golden Gate cloning site was placed downstream and upstream of the *td* intron sequence, allowing for seamless assembly of upstream and downstream exon sequences in a single one-pot reaction as described above. As we previously noticed, that the 5′ end of transcripts was preferentially acquired by the *Fs*RT–Cas1–Cas2 complex, we introduced the *td* intron within the first 23 to 31 nucleotides of the respective transcripts. We created intron-interrupted sequences of three *E. coli* genes *cspA*, *rpoS* and *argR* (which encode cold shock protein CspA, RNA polymerase sigma factor RpoS and arginine repressor, respectively). These were selected based on the fact that they were well sampled by the *Fs*RT–Cas1–Cas2 complex in preceding SENECA experiments. The flanking exon sequences were mutated in four to six positions to yield optimized sequences for *td* intron splicing, which also aided in unambiguously distinguishing the spliced and endogenous transcripts or DNA.

Accordingly, we ordered complementary oligonucleotides for the fragment of the transcript to be cloned 5′ of the *td* intron and annealed them before Golden Gate Assembly, while the fragment to be cloned 3′ of the intron was amplified by PCR from genomic DNA. Oligonucleotides were FS_1054/1055 (5′ of the intron, annealed) and FS_1056/1057 (3′ of the intron, PCR) for *CspA*; FS_1038/1039 and FS_1040/1041 for *RpoS*; FS_1046/1047 and FS_1048/1049 for *ArgR*. We ensured that mutating sequences of the respective genes to those of the *td* intron flanking sites did not generate a stop codon. The *td* intron containing *Fs*RT–Cas1–Cas2 overexpression constructs were subjected to a standard acquisition assay followed by plasmid DNA extraction, SENECA and deep sequencing. The presence of *td* intron splice sites in DNA outside of the *Fs*CRISPR array was tested by extracting gDNA from *td-ArgR* transformed cultures using the GenElute Bacterial Genomic DNA Kit (Sigma Aldrich). Libraries containing the *td* intron insertion site were amplified using a two-round PCR strategy method analogous to the ones described above using forward primers FS_1154 to FS_1157 and reverse primers FS_1158 to FS_1161 (Supplementary Table 7). First-round PCR was performed at 57 °C annealing temperature and 20 s elongation for 15 cycles. Second-round PCR was performed at 63 °C annealing temperature and 20 s elongation for 8 cycles.

Infection with MS2 phage. For infections with MS2 phage, the recording plasmid pFS_0235 was transformed into F′, and thus MS2-susceptible, NovaBlue(DE3) competent cells (EMD Millipore). The next morning, ten colonies were inoculated with 15 ml TB containing 100 μM IPTG and grown at 37 °C, 150 r.p.m. in an orbital shaker until an OD_{600} of 0.24. Then, $MgSO_4$ was added to 5 mM final concentration. Aliquots of 3 ml were split into bacterial culture tubes, infected with 200 μl high-titre MS2 phage suspension and incubated for 1 h at room temperature without shaking to allow infection by MS2. Next, culture tubes were transferred to the orbital shaker and incubated overnight at 30 °C, 80 r.p.m. Growth of *E. coli* in the presence of MS2 phage at 30 °C rather than 37 °C prevents lysis of cells by productive MS2. The next morning, shaking was increased to 150 r.p.m. Another day later (~41 h post-infection), cultures were pelleted by centrifugation, and plasmid DNA was extracted and subjected to SENECA followed by deep sequencing.

Synthetic recording of *sfGFP* and *Rluc* transcripts. The Pcat-tetR-term_PtetO encoding fragment was amplified with primers FS_1123/FS_1125 from pLP167 (kind gift from L. Pestalozzi), digested with BamHI/AgeI and cloned into AgeI/BbsI-digested pFS_0238 (see cloning of *td* intron constructs), yielding pFS_0270 which contains a BbsI-mediated Golden Gate immediately downstream of the PtetA promoter. Subsequently, *sfGFP* was amplified from pLP167 with primers FS_1134/FS_1135 and *Rluc* was amplified using FS_1136/FS_1137 from BBa_J52008 (iGEM Registry of Standard Biological Parts (http://parts.igem.org)). Both fragments were cloned into pFS_0270 using BbsI-mediated Golden Gate assembly[57], yielding pFS_0271 (*sfGFP*) and pFS_0272 (*Rluc*). LuxR promoter parts were amplified with primers FS_1584/ FS_1585 from pIG0046 and FS_1586/ FS_1587 from pIG0059 (iGEM Registry of Standard Biological Parts) and cloned into AgeI-digested pFS_0270 using NEBuilder HiFi DNA Assembly Master Mix (NEB), resulting in pFS_0399. Oligos FS_1588/ FS_1589 were annealed and cloned into pFS_0399 digested with SalI/BamHI, yielding pFS_0400. The *Fluc* coding sequence was amplified from BbaI712019 (iGEM Registry of Standard Biological Parts) using FS_1618/FS_1619, digested with BsaI and cloned into BbsI-digested

pFS_0400, resulting in pFS_0412, which was used in RNA recording experiments. For each biological replicate, 50 ml of IPTG containing TB medium was inoculated with 22 colonies of *E. coli* BL21(DE3) transformed with pFS_0271(*sfGFP*), pFS_0272 (*Rluc*) or pFS_0412(*Fluc*). When reaching an OD_{600} of 0.25, cells were split into 3-ml aliquots in bacterial culture tubes and induced with aTc in case of P_{tetA} promoter or *N*-(3-oxododecanoyl)-L-homoserine lactone (3OC6-HSL) (Sigma Aldrich) in case of P_{LuxR} promoter, and cultured in an orbital shaker for 12–14 h at 300 r.p.m., followed by plasmid DNA extraction, SENECA and deep sequencing. Spacers aligning to *sfGFP*, *Rluc* and *Fluc* were quantified as described above (see Data analysis pipeline). The detected number of unique spacers per million sequencing reads was normalized, defining the sum number of spacers per biological replicate as 100% and plotted using GraphPad Prism v7.0d. For RNA-recording with pFS_0271 and pFS_0272 RNA extraction from the same cultures was performed using the RNAsnap method[58] followed by treatment with the TURBO DNA-free kit (Thermo Scientific) using 1.5 μl TURBO DNase to minimize DNA background. Reverse transcription was performed using qScript cDNA SuperMix (Quanta Bio) with 500 ng RNA sample as a template. cDNA was diluted 1:4 and quantification was performed in 2 technical replicates by qRT–PCR using TaqMan Fast Advanced Master Mix (Life Technologies) in a Roche LightCycler 96 System. Primers and probes sequences are listed in Supplementary Table 8. Absolute copy number was calculated using standard curve method and *16S* rRNA was used as a housekeeping gene. To determine mRNA copy number corresponding to the number of cells in a single SENECA reaction (6×10^9) was calculated based on the average number of 18,700 *16s* rRNA transcripts per single *E. coli* cell[59] (BNID 102992).

Orthogonal synthetic recording. The *Rluc* coding sequence was amplified using FS_1620/FS_1137 from pFS_0272 and cloned into pFS_0399 using BbsI-mediated Golden Gate assembly[57], yielding pFS_0413. The *Fluc* coding sequence was amplified from Bba_I712019 (iGEM Registry of Standard Biological Parts) using FS_1621/FS_1619, digested with BbsI and cloned into BsaI-digested pFS_0413, resulting in pFS_0414, which was subsequently used in orthogonal synthetic recording experiments.

For each biological replicate, 33 colonies of *E. coli* BL21(DE3) transformed with pFS_0414, containing (3-Oxododecanoyl)-L-homoserine lactone (3OC6-HSL)-inducible *Fluc* and aTc-inducible *Rluc* coding sequences, were inoculated with 50 ml TB medium containing 100 μM IPTG. When the cultures reached an OD_{600} of 0.25, cells were split into 3-ml aliquots in bacterial culture tubes and induced with 75 ng/ml of anhydrotetracyclinehydrochloride (aTc) (Cayman Chemical) or 10 μM of 3OC6-HSL (Sigma Aldrich) or a combination of both and cultured in an orbital shaker for 12 h at 300 r.p.m., followed by plasmid DNA extraction, SENECA, deep sequencing and parallelized RNA extraction from the same culture, followed by reverse transcription and qRT–PCR measurements. Data were analysed as described above for recording of single synthetic transcripts.

Transcriptional response to oxidative stress. Per biological replicate, 24 colonies of *E. coli* BL21(DE3) transformed with pFS_0235 the evening before (resulting in 1 colony/1.5 ml) were inoculated with 36 ml IPTG containing TB medium and 100 μM IPTG and shaken in a 250-ml baffled shaker flask until reaching an OD_{600} of 0.24–0.25. Then cultures were split into 3-ml aliquots into bacterial culture tubes (Grainer) and treated with H_2O_2 (30% w/w solution, Sigma Aldrich) to a final concentration of 1 mM or an equal volume of ddH_2O. Growth was continued for 12 h at 300 r.p.m. followed by harvesting of 2 ml of culture for plasmid DNA extraction, SENECA and deep sequencing. Data were analysed as described above (see Data analysis pipeline).

Transcriptional response to acid stress. For pH-controlled growth, potassium-modified LB (10 g/l tryptone, 5 g/l yeast extract, 7.45 g/l KCl) was buffered with 100 mM HOMOPIPES (homopiperazine-1,4-bis(2-ethanesulfonic acid)). Subsequently, the pH of the medium was adjusted to either 5.0 (acid stress) or 7.0 (neutral) using KOH solution as described previously[35]. For each biological replicate, 33 colonies of *E. coli* BL21(DE3) transformed with pFS_0235 (resulting in 1 colony/1.5 ml) were inoculated with 50 ml of pH-adjusted, IPTG-containing LB medium. Samples were harvested at an OD_{600} of 0.3–0.6 for plasmid DNA extraction, SENECA and deep sequencing. Data were analysed as described above (see Data analysis pipeline).

Cloning of aTc-inducible *Fs*RT–Cas1–Cas2 expression construct. For recording the transcriptional response to paraquat, an aTc-inducible *Fs*RT–Cas1–Cas2 expression construct was generated. Therefore, a fragment containing the tet repressor driven by a constitutive promoter and the P_{tetA} promoter was amplified from pFS_0271 using FS_1574/1575 and digested with BglI/SphI, furthermore the N terminus of *Fs*RT–Cas1–Cas2 was amplified with FS_1576/1577 and digested with SphI/BglII. These two fragments were cloned into BglI/BglII-digested pFS_0235 yielding pFS_0393. The codon-optimized *Fs*RT–Cas1–Cas2 sequence (Supplementary Data) was obtained from Genscript, amplified using FS_1641/1642 and cloned into pFS_0393 using XhoI/SphI replacing the initial *Fs*RT–Cas1–Cas2 coding sequence and yielding pFS_0453.

Transcriptional response to 1 mM or 10 mM paraquat. Paraquat dichloride hydrate (PESTANAL, Sigma Aldrich) was dissolved at 1 M in ddH_2O. For each biological replicate, 50 colonies of *E. coli* BL21(DE3) transformed with pFS_0393 were inoculated with 75 ml TB medium containing 30 ng/ml aTc and shaken in baffled shaker flasks until they reached an OD_{600} of 0.24–0.25. Then, cultures were split into 3-ml aliquots into bacterial culture tubes and treated with either 1 mM or 10 mM paraquat and cultured for an additional 11–12 h before harvesting of 2 ml of culture for plasmid DNA extraction, SENECA and deep sequencing. Data were analysed as described above (see Data analysis pipeline).

Transcriptional response to transient paraquat exposure. For each biological replicate two colonies of *E. coli* BL21(DE3) transformed with pFS_0453 were inoculated with 3 ml of TB medium containing 30 ng/ml aTC in standard bacterial culture tubes. For the first 12 h all cultures were cultivated in the absence of paraquat (300 r.p.m., 37 °C). Then 2 ml of culture was aspirated, while the remaining 1 ml was spun down (2,300*g*, 10 min), after which the supernatant was aspirated and the bacterial pellet resuspended in 3 ml of fresh TB medium containing 30 ng/ml of aTc. For both transient and permanent stimulus conditions, paraquat was added to 10 mM final concentration and the cultures were grown for an additional 12 h as above. Then 2 ml of culture was removed, the remaining 1 ml was pelleted as above and resuspended in 3 ml of fresh TB medium containing 30 ng/ml of aTc. Paraquat was added to 10 mM for the permanent stimulus condition and cultures were grown for an additional 12 h as above. Then 2 ml of culture was harvested for plasmid DNA extraction, SENECA and deep sequencing. Additionally, 100 μl of culture was harvested for RNA extraction by the RNASnap protocol as described above followed by treatment with the TURBO DNA-free Kit (Thermo Scientific) using 1.5 μl of TURBO DNase. Ribosomal RNA was depleted using Ribo-Zero rRNA Removal Kit (Illumina) followed by library prep using TruSeq Stranded mRNA (Illumina) and deep sequencing on an NextSeq 500/550 High Output v2 kit (75 cycles) sequencing each library at a depth of 4 million reads or greater.

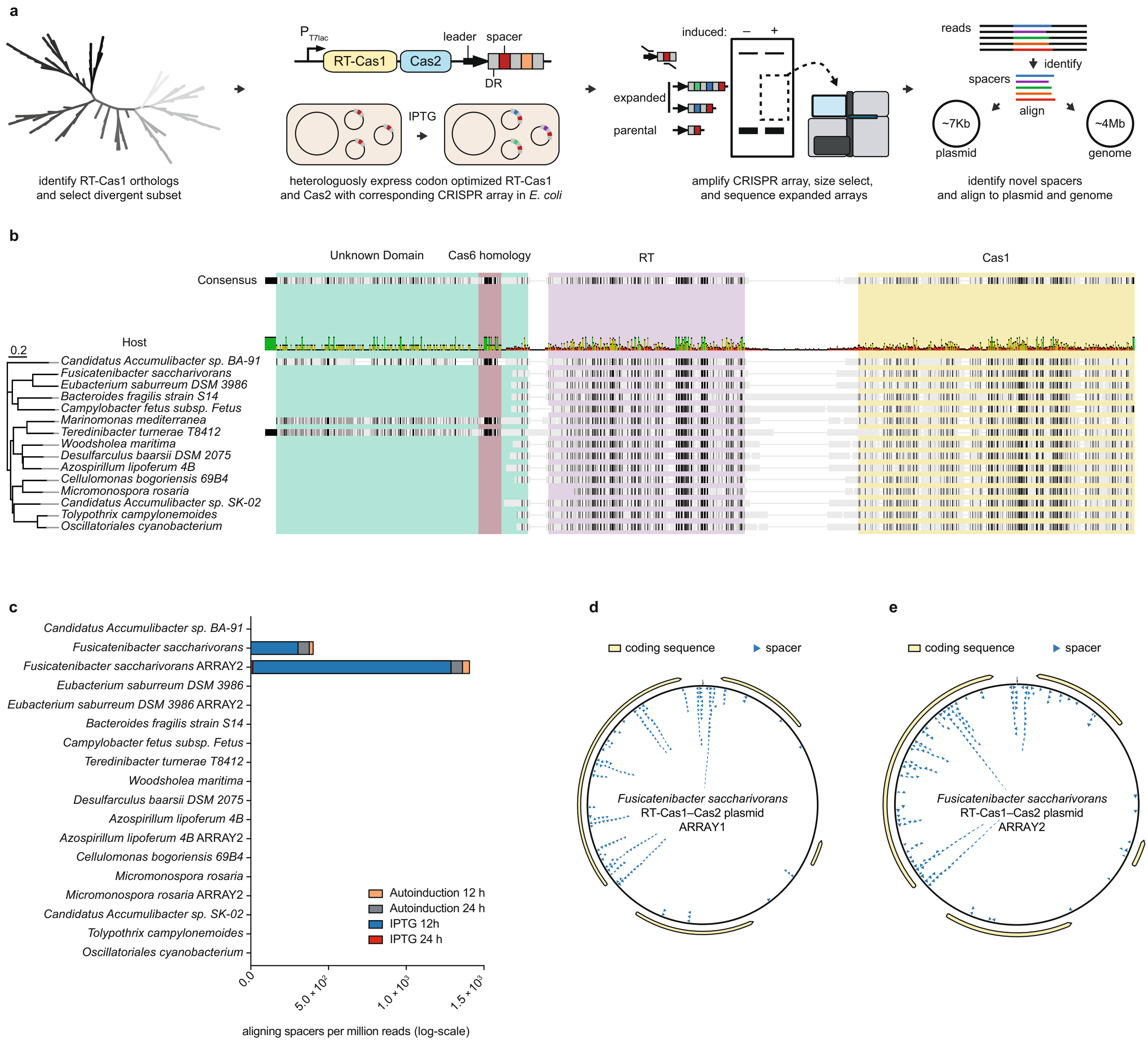
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The custom scripts used for the described data analysis are available on the Platt Laboratory website (http://www.platt.ethz.ch).

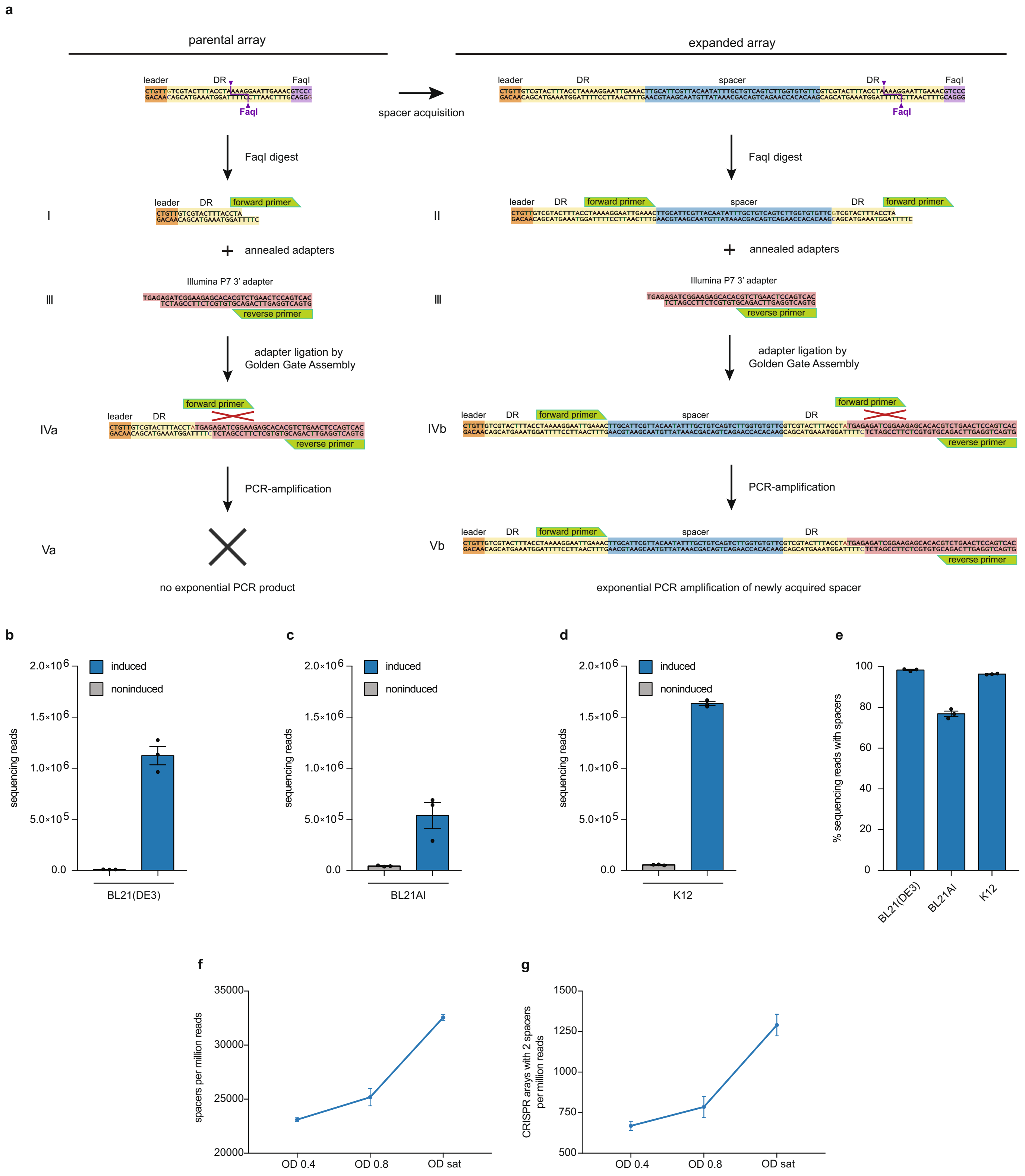
Data availability

Deep sequencing data are available in the National Center for Biotechnology Information Sequence Read Archive (PRJNA484149). The data sets generated and/or analysed during the current study are available from the corresponding author upon reasonable request.

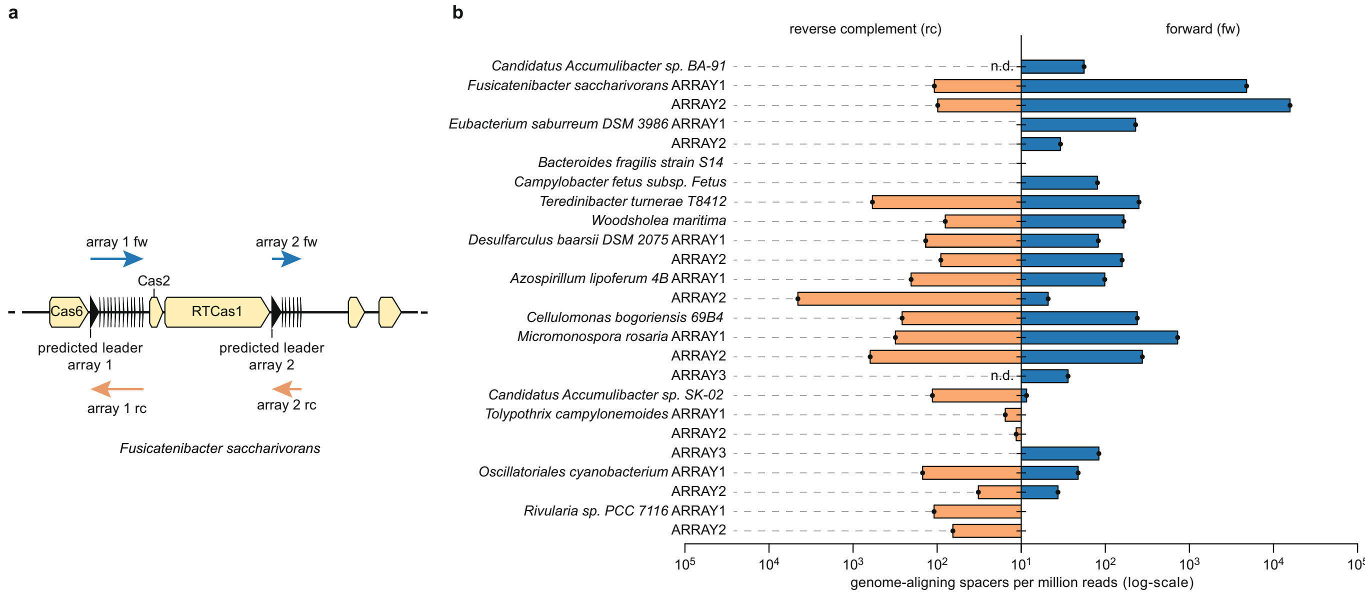
46. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
47. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).
48. Zhang, Q. & Ye, Y. Not all predicted CRISPR–Cas systems are equal: isolated cas genes and classes of CRISPR like elements. *BMC Bioinformatics* **18**, 92 (2017).
49. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
50. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
54. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
55. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
56. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
57. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS ONE* **4**, e5553 (2009).
58. Stead, M. B. et al. RNAsnap™: a rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. *Nucleic Acids Res.* **40**, e156 (2012).
59. Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res.* **38**, D750–D753 (2010).



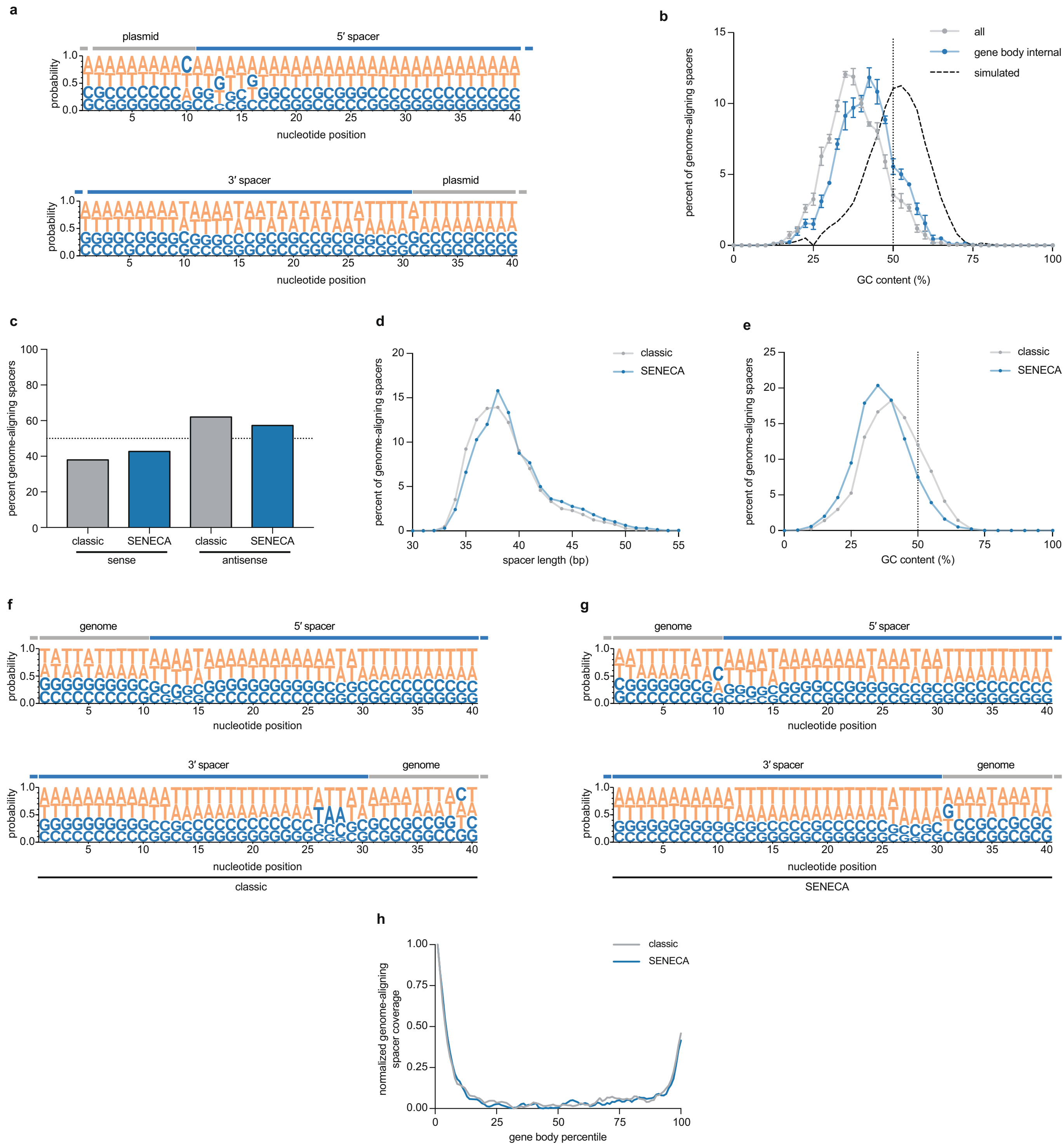
Extended Data Fig. 1 | RT–Cas1 orthologue search and screening. **a**, Experimental workflow involving the identification of 121 RT–Cas1 orthologues, overexpression in *E. coli* from the plasmid carrying minimal CRISPR array, containing leader–DR–spacer1–DR–spacer2–DR, followed by deep sequencing of expanded CRISPR arrays, and analysis and characterization of identified spacers. **b**, A comparison of the 14 disparate RT–Cas1 proteins selected for functional testing. Indicated on the left is the host species followed by a neighbour-joining phylogenetic tree built using Jukes–Cantor genetic distances of a MUSCLE multiple sequence alignment. The large 'unknown domain' is highlighted in green, Cas6 homology domain in pink, RT domain in purple, and Cas1 in yellow. **c**, Detection frequency of newly acquired spacers after overnight growth and induction of RT–Cas1–Cas2 in *E. coli* BL21(DE3) in different induction media. Shown is the sum of spacer counts per 1 million sequencing reads, $n = 1$ biological sample. **d**, Representative alignments of 200 spacers sequenced from *F. saccharivorans* array 1 to the corresponding overexpression plasmid. **e**, Representative alignments of 200 spacers sequenced from *F. saccharivorans* array 2 to the corresponding overexpression plasmid.



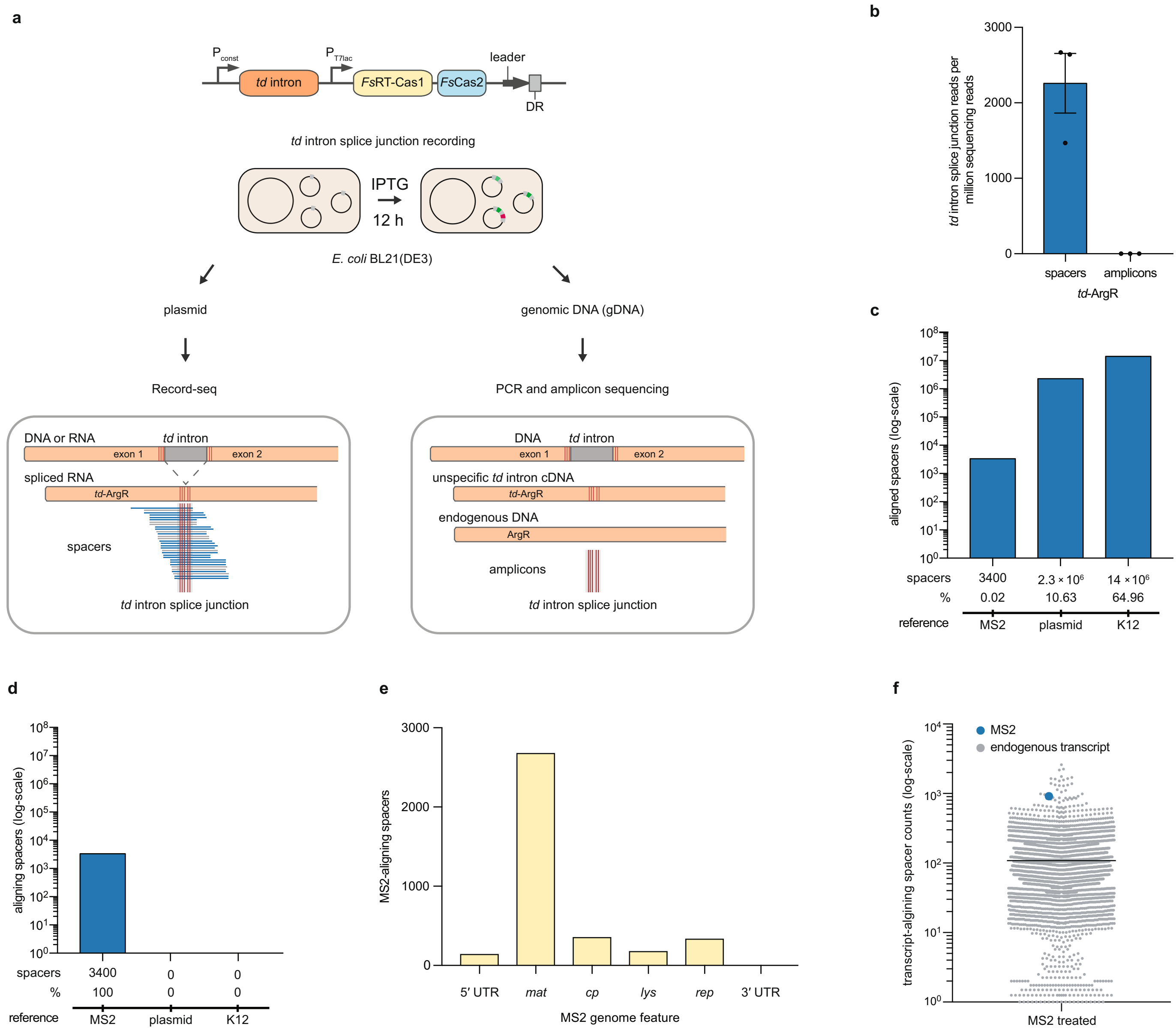
Extended Data Fig. 2 | SENECA workflow and assessment of Record-seq efficiency in different culture conditions. **a**, SENECA relies on a plasmid containing a minimal CRISPR array consisting of the leader sequence followed by a single direct repeat and a recognition sequence for the restriction enzyme FaqI. The SENECA workflows for the parental (left) and expanded (right) arrays are shown. In a Golden Gate reaction, FaqI cleaves within the direct repeat (I/II), introducing sticky ends for ligation to an Illumina P7 3′ adaptor (III). For the parental array this results in a single truncated direct repeat (IVa). For the expanded array this results in a truncated direct repeat as well as an intact direct repeat and spacer (IVb). PCR with primers binding to the full-length direct repeat and the Illumina P7 3′ adaptor results in linear amplification of the parental array (Va) and exponential amplification of the expanded array (Vb). **b**, Sequencing reads obtained from *E. coli* BL21(DE3) cells transformed with *Fs*RT–Cas1–Cas2-encoding plasmid with or without IPTG induction. **c**, As in **b** but in *E. coli* BL21AI. **d**, As in **b** but in *E. coli* NovaBlue(DE3), a K12 substrain of *E. coli*. **e**, Percentage of sequencing reads from induced samples containing newly acquired spacers. **f**, Spacers per million sequencing reads obtained from cultures at an OD_{600} of 0.4, 0.8 or upon saturation. **g**, CRISPR arrays with two spacers per million sequencing reads obtained from cultures at an OD_{600} of 0.4, 0.8 or upon saturation. Values in **b**–**g** are mean ± s.e.m., $n = 3$ independent biological samples.



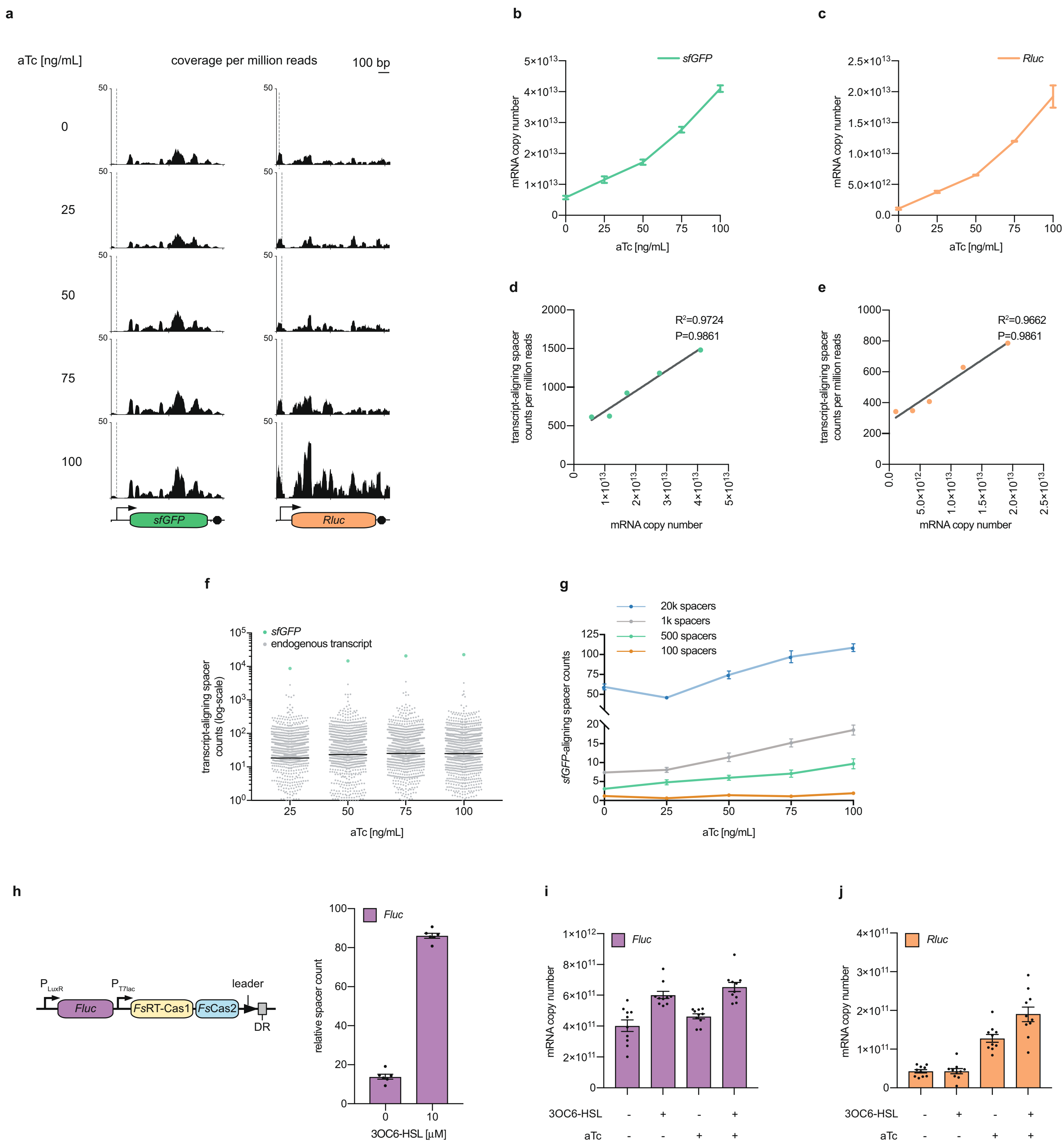
Extended Data Fig. 3 | Record–seq-based screen of RT–Cas1 orthologues and CRISPR array directionalities. a, Schematic of the *F. saccharivorans* CRISPR locus depicting the selection of CRISPR arrays and directionalities for Record–seq analysis. CRISPR arrays within each locus were identified and cloned into plasmids encoding corresponding RT–Cas1–Cas2 coding sequences. Arrays were tested in both possible directionalities, forward and reverse with a 150-bp leader. In cases of insufficient genomic data, arrays were tested in only one directionality. **b**, Record–seq readout of RT–Cas1 orthologues and CRISPR array directionalities. Acquisition efficiencies for forward (fw) and reverse complement (rc) directionality of each array are plotted in blue and orange, respectively. Values are genome-aligning spacers per million sequencing reads, $n = 1$ biological sample. n.d., not determined.



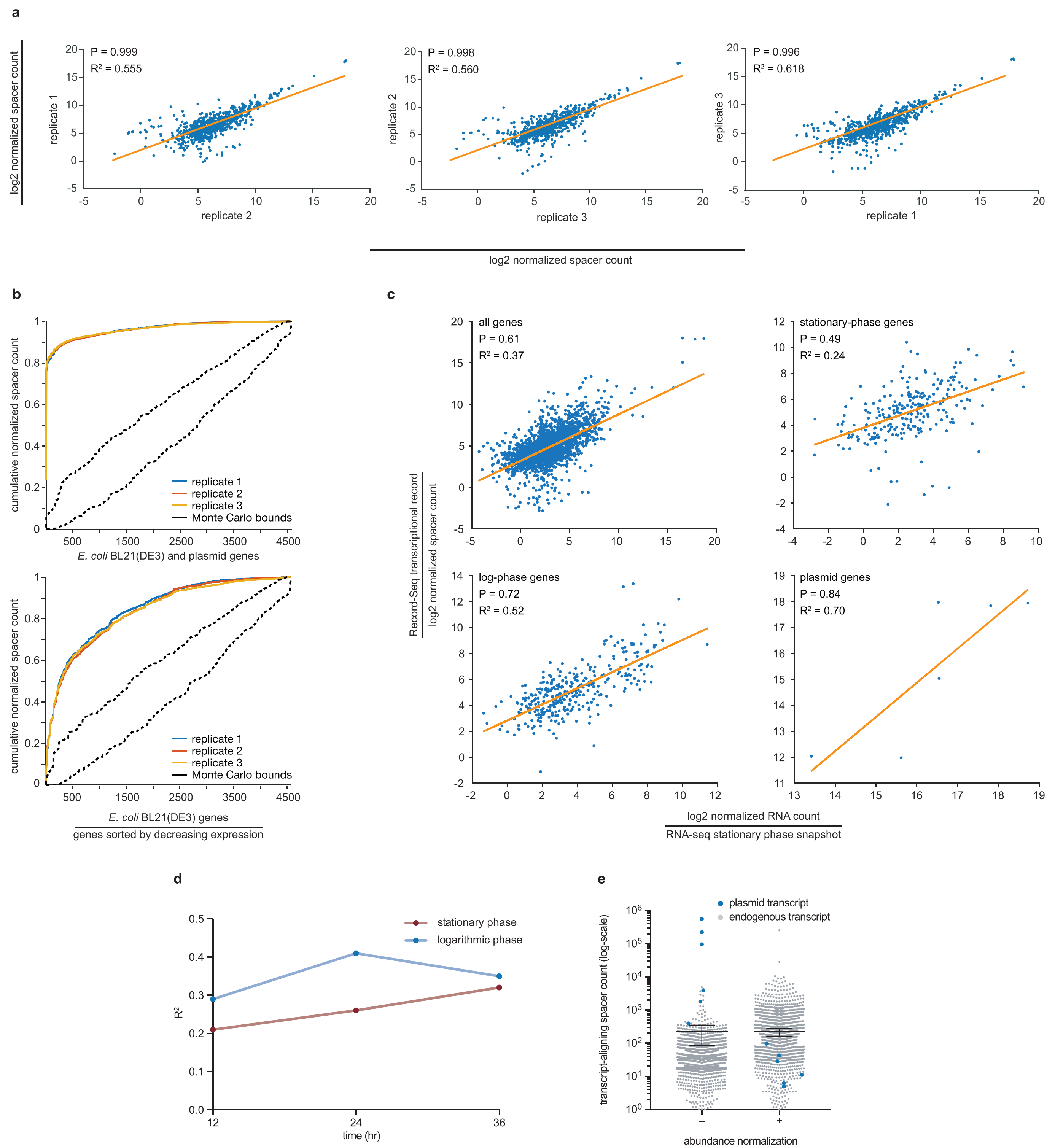
Extended Data Fig. 4 | Characterization of spacers acquired by *Fs*RT–Cas1–Cas2 and comparison of SENECA and classic spacer acquisition readouts. a, Nucleotide probabilities determined using plasmid-aligning spacers merged across $n = 14$ independent biological samples, prepared as for Fig. 2f. **b**, Histogram of spacer GC content for all spacers or spacers acquired internal to the body of the transcript ('gene body internal'). Values represent mean percentage of genome-aligning spacers ± s.e.m., $n = 3$ independent biological samples. **c**, Percentage of spacers aligning to either the sense or antisense strand of coding genes. The sense or antisense orientation label is with respect to the RNA, prepared as for Fig. 2c. **d**, Length distribution of genome-aligning spacers, prepared analogous to Fig. 2d. **e**, GC content distribution of genome-aligning spacers. The dotted line represents a balanced (50%) GC content, prepared as for Fig. 2e. **f**, Nucleotide probabilities for classic acquisition readout, prepared as for Fig. 2f. **g**, Nucleotide probabilities for SENECA acquisition readout, prepared analogous to Fig. 2f. Gene body coverage. For each gene the spacer coverage was determined and transformed into percentiles for comparison. Values are mean normalized coverage. $n = 1$ pooled sample, containing 5,798 spacers. Values in **c**–**g** are mean percentage of genome-aligning spacers, $n = 1$ pooled sample, containing 5,798 spacers.



Extended Data Fig. 5 | Characterization of spacers acquired by *Fs*RT–Cas1–Cas2. a, Experimental workflow for determining the specificity of *Fs*RT–Cas1–Cas2 for RNA using the *td* intron splice junction to detect RNA-derived spacers. Genomic DNA (gDNA) was extracted from an independent culture and subjected to targeted deep sequencing of the *td* intron insertion site. **b**, Quantification of *td* intron splice junctions. The splice junction is specific to RNA-derived spacers and not genomic DNA or cDNA copies generated by alternative reverse transcriptases in the *E. coli* genome. Values represent mean *td* intron splice junction counts per million sequencing reads ± s.e.m., $n = 3$ independent biological samples. **c**, Number of spacers aligned to plasmid, *E. coli* genome, and MS2 genome, showing CRISPR acquisition from an RNA virus. The total number and percentage of spacers aligning to each reference are shown. Values represent the sum of MS2-aligning spacers across replicates, $n = 64$ technical replicates from $n = 2$ biological samples, representing 22 million spacers. **d**, Number of MS2-aligned spacers from **c** that align to the overexpression plasmid, *E. coli* and MS2 genome, showing that MS2-aligned spacers are specific to the MS2 genome. The total number and percentage of MS2-aligned spacers that subsequently align to each reference are shown, $n = 64$ technical replicates from $n = 2$ biological samples, representing 22 million spacers. **e**, Total number of spacers aligning to features of the MS2 genome, $n = 64$ technical replicates from $n = 2$ biological samples, representing 22 million spacers. **f**, Scatter plot of transcript counts from the MS2 and *E. coli* genomes. Each dot represents the mean spacer count for each transcript, $n = 4$ independent biological samples. The horizontal black bars are mean genome-aligning spacer count across all transcripts ± s.e.m.

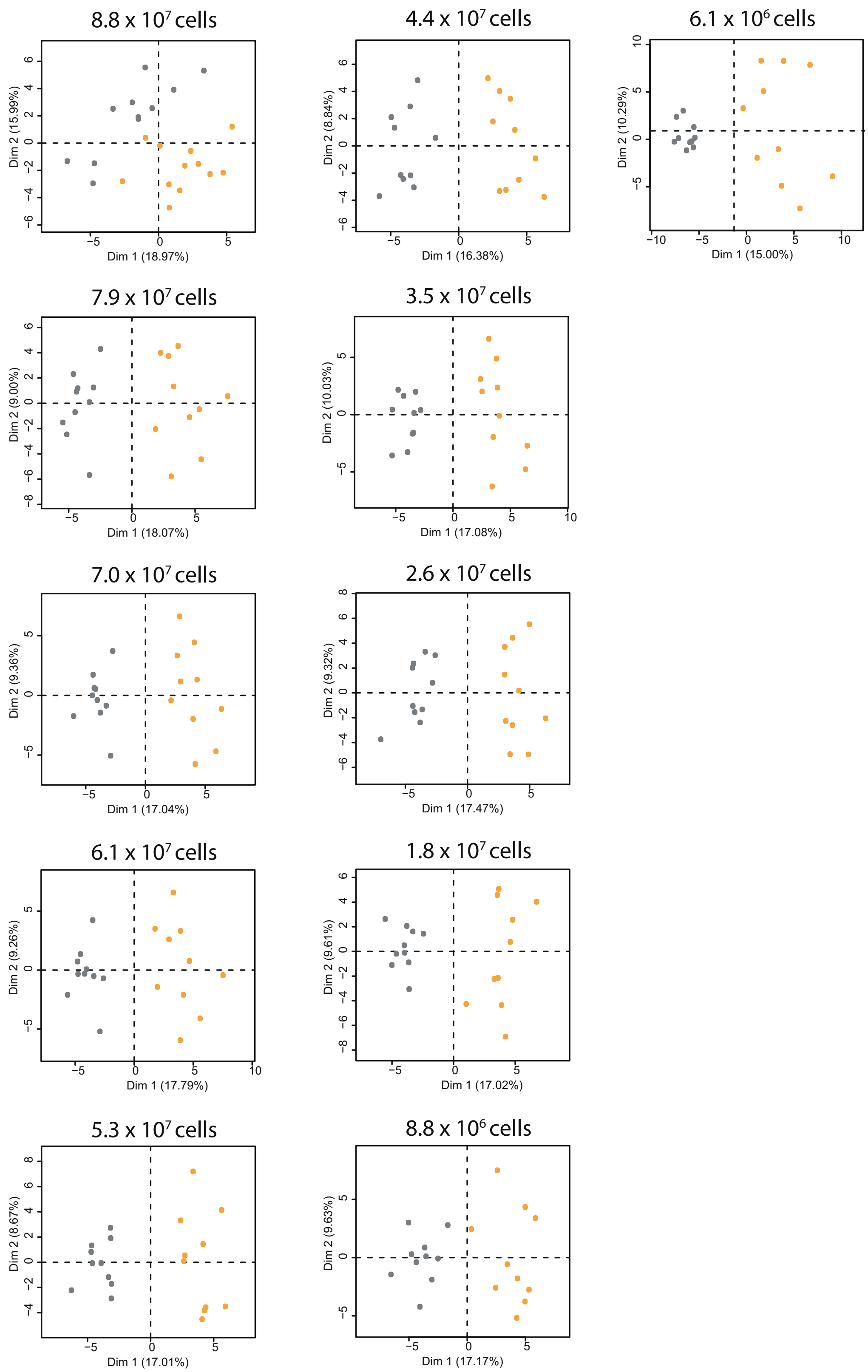


Extended Data Fig. 6 | Quantitative analysis of arbitrary RNA sequence recording using qRT–PCR and Record–seq. a, Coverage of spacers from Fig. 3f aligning to *sfGFP* or *Rluc*. Arrow and dotted line reflect the transcription start site (TSS), black octagon indicates the transcriptional terminator. For each nucleotide position, the sum spacer coverage per million sequencing reads is shown, $n = 10$ independent biological samples. **b**, Absolute quantification of *sfGFP* mRNA measured by qRT–PCR. Samples from Fig. 3f. Values are mean ± s.e.m. copy number per 6×10^9 cells, normalized by *16S* rRNA copy number, $n = 10$ independent biological samples. **c**, As in **b**, but for *Rluc*. **d**, Scatter plot depicting the correlation between absolute *sfGFP* mRNA copy number and the number of transcript-aligning spacers from Fig. 3f. Linear regression fit, coefficient of determination (R^2), and Pearson linear correlation coefficient (P), $n = 10$ independent biological samples. **e**, As in **d**, but for *Rluc*. **f**, Comparison of spacer counts for arbitrary *sfGFP* sequence and endogenous transcripts. Each dot represents the mean spacer count for each transcript, horizontal black bars are mean genome-aligning spacer count ± s.e.m., $n = 10$ independent biological samples. **g**, Dose–response relationship between *sfGFP*-aligning spacers and inducer concentration for different numbers of recorded spacers. These data represent the average number of *sfGFP*-aligning spacers ± s.e.m., $n = 10$ independent biological samples. **h**, Relative spacer count of spacers mapping to the *Fluc* transcript after 3OC6-HSL induction. Values are the normalized mean number of spacers per million sequencing reads ± s.e.m. with $n = 6$ independent biological samples. **i**, Absolute quantification of *Fluc* mRNA measured by qRT–PCR. Data were obtained from the same bacterial cultures as in Fig. 3g. Values are mean copy number per 6×10^9 cells, normalized by *16S* rRNA copy number, ± s.e.m., $n = 10$ independent biological samples. **j**, As in **i**, but for *Rluc*.

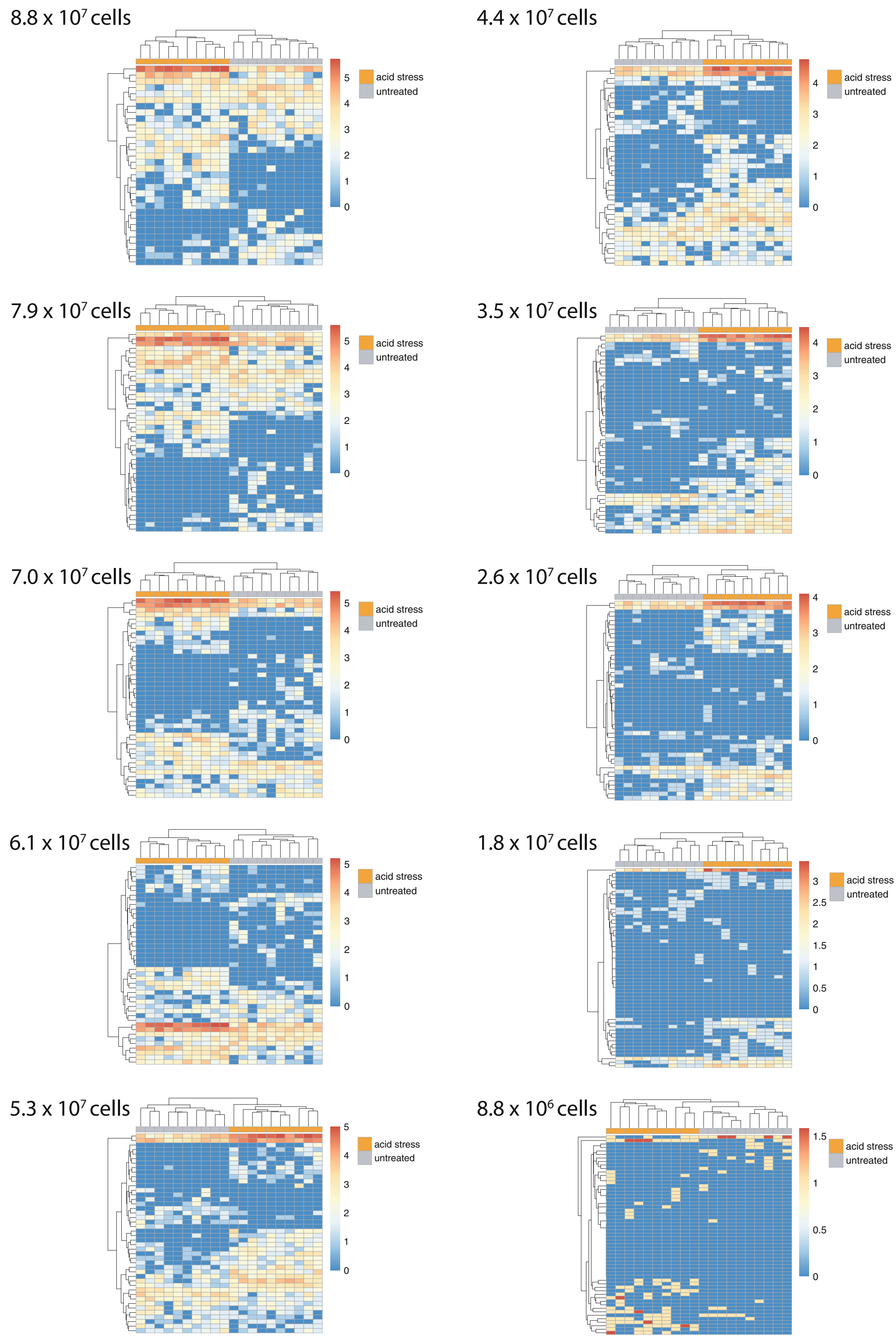


Extended Data Fig. 7 | See next page for caption.

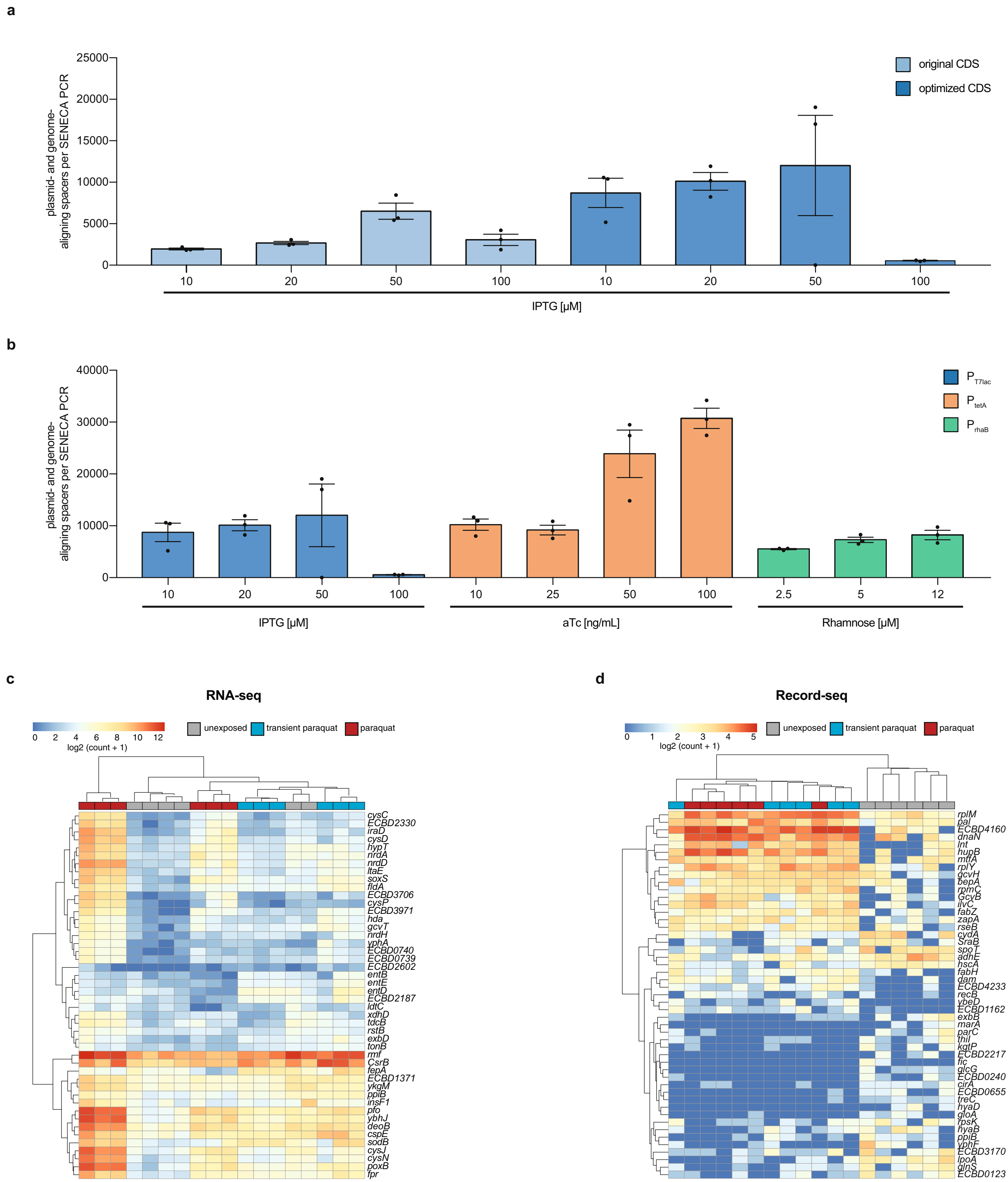
Extended Data Fig. 7 | Record–seq reveals cumulatively highly expressed genes. a, Scatter plots depicting Record–seq correlation between $n = 3$ independent biological replicates shown in **b** and **c**. Linear regression fit, coefficient of determination (R^2), and Pearson linear correlation coefficient (P) are shown for each comparison. Data represent $\log_2$-normalized transcript quantification counts. **b**, Spacers are preferentially acquired from highly expressed genes. Record–seq spacer counts for plasmid and *E. coli* genes (top) or only *E. coli* genes (bottom) according to decreasing RNA-seq-based gene expression values. Monte Carlo bounds reflect simulated spacers with no transcriptional bias. Mean cumulative normalized spacer count, and Monte Carlo bounds are shown, $n = 3$ independent biological samples. **c**, Assessing the correlation between an RNA-seq stationary phase snapshot and a Record–seq transcriptional record. RNA-seq and Record–seq were performed on the same population of *E. coli* BL21(DE3) in stationary phase growth, induced to express *Fs*RT–Cas1–Cas2 overnight. The correlation between all (top left), stationary-phase (top right), log-phase (bottom left), and plasmid-borne (bottom right) genes are shown. The linear regression fit, coefficient of determination (R^2), and Pearson linear correlation coefficient (P) are shown for each comparison. The data represent the $\log_2$ normalized transcript quantification counts averaged across replicates, $n = 3$ independent biological samples. **d**, Correlation of Record–seq with log- and stationary-phase genes over long-term cultivation. These data represent the R^2 value calculated as described for **b** for either stationary or logarithmic phase gene sets using different *E. coli* culture time points as inputs with $n = 3$ independent biological samples. **e**, Comparison of transcript-aligning spacer counts with and without normalizing for gene expression level. Each dot represents the mean normalized number of counts per transcript with $n = 3$ independent biological samples. The horizontal black bars are mean genome-aligning spacer count ± s.e.m. See Supplementary Notes for detailed discussions on **b**, **d**.



Extended Data Fig. 8 | Defining the minimum number of cells required for assessing complex cellular behaviours using Record–seq and PCA. Using the acid stress response data set shown in Fig. 4, PCA was performed on the entire data set as well as progressively and randomly downsampled data. These data show that Record–seq appropriately classifies the acid stress response samples with 7% of the original data (corresponding to 314 spacer or 6.1×10^6 *E. coli* cells). The calculation of the required number of *E. coli* cells is described in detail in the Supplementary Notes; $n = 10$ independent biological samples.



Extended Data Fig. 9 | Defining the minimum number of cells required for assessing complex cellular behaviours using Record–seq and differential expressed signature gene analysis. Using the acid stress response data set shown in Fig. 4e–g, differential expressed signature genes were identified for the entire data set as well as progressively and randomly downsampled data. The plots depict hierarchically clustered signature gene heatmaps. These data show that with 10% of the original data (corresponding to 448 spacer or 8.8×10^6 *E. coli* cells) the signature genes can appropriately classify the samples. The calculation of cell numbers is described in detail in the Supplementary Notes; $n = 10$ independent biological samples.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Optimization of CRISPR spacer acquisition efficiency and detection of signature genes corresponding to Record–seq-compatible sentinel cells for encoding transient herbicide exposure. **a**, Plasmid and genome-aligning spacers obtained from *E. coli* BL21(DE3) transformed with *Fs*RT–Cas1–Cas2 encoding plasmid using the original coding sequence (CDS) (light blue) or optimized CDS (dark blue) under the indicated IPTG concentrations. **b**, Plasmid and genome-aligning spacers obtained from *E.coli* BL21(DE3) transformed with *Fs*RT–Cas1-Cas2 encoding plasmid using the optimized coding sequence under transcriptional control of the P_{T7lac}, P_{tetA}, or P_{rhaB} promoter, induced with the indicated concentrations of IPTG, aTc, or Rhamnose, respectively. **c**, Unsupervised hierarchical clustering of RNA-seq cumulative expression profiles for signature differentially (cumulatively) expressed genes. Signature genes represent the union between the top 20 most differently expressed genes identified by DESeq2, edgeR, and baySeq, $n = 6$ independent biological samples. **d**, Unsupervised hierarchical clustering of Record–seq cumulative expression profiles for signature differentially (cumulatively) expressed genes. Signature genes represent the union between the top 20 most differently expressed genes identified by DESeq2, edgeR, and baySeq, $n = 6$ independent biological samples. Data in **a**, **b** are mean ± s.e.m., $n = 3$ independent biological samples.

https://doi.org/10.1038/s41586-018-0588-y

The dispersion–brightness relation for fast radio bursts from a wide-field survey

R. M. Shannon[1,2,3,4]*, J.-P. Macquart[3,5]*, K. W. Bannister[4], R. D. Ekers[3,4], C. W. James[3,5], S. Osłowski[1], H. Qiu[4,5,6], M. Sammons[3], A. W. Hotan[7], M. A. Voronkov[4], R. J. Beresford[4], M. Brothers[4], A. J. Brown[4], J. D. Bunton[4], A. P. Chippendale[4], C. Haskins[7], M. Leach[4], M. Marquarding[4], D. McConnell[4], M. A. Pilawa[4], E. M. Sadler[5,6], E. R. Troup[4], J. Tuthill[4], M. T. Whiting[4], J. R. Allison[8], C. S. Anderson[7], M. E. Bell[4,5,9], J. D. Collier[4,10], G. Gürkan[7], G. Heald[7] & C. J. Riseley[7]

Despite considerable efforts over the past decade, only 34 fast radio bursts—intense bursts of radio emission from beyond our Galaxy—have been reported[1,2]. Attempts to understand the population as a whole have been hindered by the highly heterogeneous nature of the searches, which have been conducted with telescopes of different sensitivities, at a range of radio frequencies, and in environments corrupted by different levels of radio-frequency interference from human activity. Searches have been further complicated by uncertain burst positions and brightnesses—a consequence of the transient nature of the sources and the poor angular resolution of the detecting instruments. The discovery of repeating bursts from one source[3], and its subsequent localization[4] to a dwarf galaxy at a distance of 3.7 billion light years, confirmed that the population of fast radio bursts is located at cosmological distances. However, the nature of the emission remains elusive. Here we report a well controlled, wide-field radio survey for these bursts. We found 20, none of which repeated during follow-up observations between 185–1,097 hours after the initial detections. The sample includes both the nearest and the most energetic bursts detected so far. The survey demonstrates that there is a relationship between burst dispersion and brightness and that the high-fluence bursts are the nearby analogues of the more distant events found in higher-sensitivity, narrower-field surveys[5].

Since the beginning of 2017, we have been surveying for fast radio bursts (FRBs) using a subset of the Australian Square Kilometre Array Pathfinder[6] (ASKAP), a radio-telescope array comprising 36 antennas that is currently being commissioned. Each antenna is equipped with a phased-array-feed (PAF) receiver that is sensitive to 30 deg^2 on the sky at the prime focus of a 12-metre reflector. The searches, conducted at a central frequency of 1.3 GHz, have used a fly's-eye configuration, with each of 5−12 available antennas pointed towards a different area of sky, widening our field of view to target the brightest portion of the burst population. The system set-up and search algorithms are identical to those reported previously from ASKAP[7]. The PAFs allow full sampling of the focal plane, making it possible to measure the burst positions with as little as 10×10 arcmin uncertainty[7] (90% confidence), and the fluence of each burst to better than 20% accuracy[7]—in contrast with previous searches where burst positions were unconstrained within the antenna beam pattern, and burst fluences could be uncertain by factors greater 10. The searches have been conducted at 57 high Galactic latitude pointings, $|b| = 50 \pm 5$ deg, removing the need to account for potential bias in the latitude dependence of the rate[8] and minimizing the contribution of the Galaxy to the electron column density. Observations of these pointings were interleaved with short scans of known pulsars to check system performance (see Supplementary Information, section 1). Pointings were revisited a median of 570 times over the course of the survey. In total, our survey exposure is 5.1×10^5 deg^2 h (see Supplementary Information, section 1).

We have discovered 20 FRBs in total (including FRB 170107, reported previously[7]), the properties of which are listed in Table 1. The electron column densities towards the bursts, expressed as dispersion measures, range from 114 pc cm^{-3} to 992 pc cm^{-3}. The dispersion measures contain contributions from the Milky Way, the host galaxy, and the intergalactic medium. The Galactic dispersion-measure contributions are likely to be less than 60 pc cm^{-3} for all of our bursts. Assuming a negligible circumburst environment, the host galaxy is likely to have a similar median contribution[9] (see Supplementary Information, section 4). Larger contributions would be possible if the bursts were embedded in dense nebulae such as supernova remnants or in galaxy centres. However, our lowest dispersion-measure events show that for some objects this contribution does not exceed 50−120 pc cm^{-3}. We also assume that the intergalactic medium is uniformly distributed, and use a standard distance/dispersion-measure relationship[10] to infer distances from the extragalactic dispersion-measure component. For the nearest (and hence lowest dispersion measure) bursts, larger stochasticity in the intergalactic dispersion-measure contribution would be expected, as it will depend strongly on the host location within the corresponding local large-scale structure[11], which has a characteristic size scale of 30 Mpc. For more distant objects, bursts will propagate through many voids and walls in the large-scale structure and therefore will only have modest 10−20% scatter[11]. Our sample includes the lowest hitherto reported dispersion measure for a burst (FRB 171020), which has a column-density excess beyond the Milky Way[12] of 80 pc cm^{-3}. Given the assumed distance/dispersion-measure model described above, this burst would have originated at a distance of approximately 130 Mpc (at a redshift, z, of around 0.03). This burst has one of the poorest localizations (30×50 arcmin ellipse, 90% confidence) in our sample, because it was detected in a corner beam. However, there is only one galaxy catalogued at a distance of less than 210 Mpc (z less than 0.05) within the uncertainty region of the FRB. This is the distorted Sc-type spiral galaxy PGC 068417, at a distance of 37 Mpc[13] ($z = 0.0087$). A second galaxy at the edge of the error box, PGC 3094828, has a catalogued redshift of $z = 0.0665$, placing it at a distance of 275 Mpc (ref. [14]). Low-dispersion bursts such as FRB 171020 offer the potential for detailed host-galaxy studies.

The measured fluences range from 34 Jy ms to 420 Jy ms, with the latter being the highest well constrained fluence (see Supplementary Information, section 4). Above our completeness threshold of a signal-to-noise ratio of 9.5—which corresponds to a fluence of 26 Jy ms $(w/1.26\ \mathrm{ms})^{-1/2}$ for a pulse duration w matching our time resolution of 1.26 ms—the event rate is 37 ± 8 per day over the entire sky

[1]Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Victoria, Australia. [2]ARC Centre of Excellence for Gravitational Wave Discovery (OzGrav), Hawthorn, Australia. [3]International Centre for Radio Astronomy Research, Curtin Institute of Radio Astronomy, Curtin University, Perth, Western Australia, Australia. [4]CSIRO Astronomy and Space Science, Australia Telescope National Facility, Epping, New South Wales, Australia. [5]ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO), Sydney, Australia. [6]Sydney Institute for Astronomy, School of Physics, University of Sydney, Sydney, New South Wales, Australia. [7]CSIRO Astronomy and Space Science, Australia Telescope National Facility, Bentley, Western Australia, Australia. [8]Sub-Department of Astrophysics, Department of Physics, University of Oxford, Oxford, UK. [9]School of Mathematics and Physical Sciences, University of Technology Sydney, Sydney, New South Wales, Australia. [10]School of Computing, Engineering, and Mathematics, Western Sydney University, Sydney, New South Wales, Australia. *e-mail: rshannon@swin.edu.au; J.Macquart@curtin.edu.au

Table 1 | Properties of ASKAP FRBs

FRB	Time (TAI*)	DM (pc cm^{-3})	E_ν (Jy ms)	R.A. (hh:mm)†	Dec. (dd:mm)†	g_l (deg.)	g_b (deg.)	w (ms)	S/N‡	T_{obs} (d)	$T_{\pm 15}$ (h)
170107	20:05:45.1393(1)	609.5(5)	58(3)	11:23.3(7)	−05:00(10)	266.0	54.1	2.4(2)	16.0	27.9	15
170416	23:11:49.7994(2)	523.2(2)	97(2)	22:13(1)	−55:02(9)	337.6	−50.0	5.0(6)	13.0	16.2	31
170428	18:03:11.7003(2)	991.7(9)	34(2)	21:47(2)	−41:51(20)	359.2	−49.9	4.4(5)	10.5	31.6	22
170707	06:18:11.3548(2)	235.2(6)	52(3)	02:59(2)	−57:16(20)	269.1	−50.5	3.5(5)	9.5	11.3	56
170712	13:22:54.39488(8)	312.79(7)	53(2)	22:36(1)	−60:57(10)	329.3	−51.6	1.4(3)	12.7	7.7	32
170906	13:07:33.48832(8)	390.3(4)	74(7)	21:59.8(4)	−19:57(10)	34.2	−49.5	2.5(3)	17.0	32.7	95
171003	04:08:00.78117(9)	463.2(1.2)	81(5)	12:29.5(7)	−14:07(20)	283.4	46.3	2.0(2)	13.8	27.3	67
171004	03:24:16.2501(1)	304.0(3)	44(2)	11:57.6(8)	−11:54(10)	282.2	48.9	2.0(3)	10.9	30.5	84
171019	13:27:17.09738(1)	460.8(1.1)	219(5)	22:17.5(5)	−08:40(7)	52.5	−49.3	5.4(3)	23.4	17.6	57
171020¶	10:28:35.59870(4)	114.1(2)	200^{+500}_{-100}	22:15(3)	−19:40(40)	29.3	−51.3	1.7(2)	19.5	32.7	95
171116	15:00:10.3052(2)	618.5(5)	63(2)	03:31.0(6)	−17:14(10)	205.0	−49.8	3.2(5)	11.8	45.7	102
171213	14:23:17.46705(3)	158.6(2)	133(12)	03:39(2)	−10:56(20)	200.6	−48.3	1.5(2)	25.1	35.5	118
171216	17:59:47.82229(9)	203.1(5)	40(2)	03:28(1)	−57:04(10)	273.9	−48.4	1.9(3)	8.0	7.7	26
180110‖	07:35:11.9590(1)	715.7(2)	420(20)	21:53.0(7)	−35:27(7)	7.8	−51.9	3.2(2)	35.6	37.7	154
180119‖	12:25:07.7476(2)	402.7(7)	110(3)	03:29.3(5)	−12:44(8)	199.5	−50.4	2.7(5)	15.9	35.5	125
180128.0	01:00:15.6179(1)	441.4(2)	51(2)	13:56(1)	−06:43(20)	326.7	52.2	2.9(3)	12.4	21.6	102
180128.2	04:54:03.7962(1)	495.9(7)	66(4)	22:22(2)	−60:15(11)	327.8	−48.6	2.3(2)	9.6	11.3	62
180130‖	04:56:06.9932(1)	343.5(4)	95(3)	21:52.2(9)	−38:34(10)	5.9	−51.8	4.1(1.0)	10.3	37.7	120
180131	05:45:42.3207(2)	657.7(5)	100(3)	21:49.9(8)	−40:41(8)	0.6	−50.7	4.5(4)	13.8	31.6	103
180212	23:45:41.39991(9)	167.5(5)	96(8)	14:21(2)	−03:35(30)	338.3	50.0	1.81(6)	18.3	21.9	79

Uncertainties are listed in parentheses. Dec., declination; DM, dispersion measure; g_l, Galactic longitude of beam centre; g_b, Galactic latitude of beam centre; R.A., right ascension; T_{obs}, total observation time on-field; $T_{\pm15}$, time on-field within ± 15 days of FRB; w, burst full width at half maximum intensity.

*Burst arrival time referenced to a frequency of 1,297.5 MHz relative to the TAI time standard.

†Position uncertainties are 90% confidence limits and referred to the epoch J2000. Posterior localization regions are reproduced in the Supplementary Figures and are available as supplementary files, as noted in the Data availability statement.

‡S/N is the signal-to noise ratio reported in the primary beam by the search algorithm.

¶Quoted errors on fluence for FRB 171020 represent a 90% confidence limit.

‖Pulse widths are reported after deconvolution of exponential pulse broadening function (see Supplementary Information, section 1).

(1σ uncertainty). This is a factor of 200 less than the rate obtained from high Galactic latitude searches with the 64-metre Parkes radio telescope[5]. This difference is a consequence of several factors, including the burst fluence and width distributions, telescope sensitivities, and instrumental selection effects.

All bursts in the sample show strong spectral modulation, as seen in Fig. 1. While some, such as FRB 170712, show broad-band structure that persists across half of the band, many exhibit power concentrated in narrower structures of a few megahertz bandwidth, with signals absent in large fractions of the band. Others, such as FRB 170906, show strong narrow-band features imposed on broad-band structures. It is unclear whether the spectral modulation is intrinsic to the source, caused by diffractive scintillation, or a consequence of refractive propagation effects such as caustic-induced magnification[15]. The Parkes population[5]—comprising 26 bursts detected in the same frequency band as ASKAP—shows less evidence for strong spectral modulation, even at high Galactic latitudes, where Galactic propagation would be expected to impart similar spectral structure to the ASKAP bursts. There are notable counterexamples in the Parkes sample that do show spectral modulation: the highest signal-to-noise ratio Parkes detection FRB 180309 (ref. [16]); and FRB 150807, which shows spectral modulation on two scales[17]. Studies of the bright population in a different frequency range would distinguish between causes for the spectral modulation, because propagation effects are strongly frequency dependent. We averaged the spectra to estimate the global properties of the detections, finding the spectral index of combined spectra of all the bursts to be steep, with $\beta = -1.8 \pm 0.3$ over our observing band (where fluence, E, scales proportionally to frequency ν as $E(\nu) \propto \nu^{\beta}$) (see Supplementary Information, section 2).

Temporal analyses of the burst profiles are limited by the 1.26-ms time resolution of our present datasets. All of the bursts are marginally resolved beyond instrumental dispersion smearing, with a median burst full width at half maximum being 3.0 ms. This is comparable to the median burst width of the Parkes sample[18]. Three bursts (FRBs 180110, 180119 and 180130) show exponential profiles consistent with scatter broadening. The broadening time for FRB 180110 varies with frequency, scaling proportionally to $\nu^{-3.5 \pm 0.6}$, consistent with propagation through turbulent plasma[19]. FRBs 180119 and 180130 are too weak to allow us to conclusively detect a change in pulse broadening with frequency. All three also show spectral modulation; if this is caused by scattering, the presence of both pulse broadening and spectral modulation indicates that there is propagation through two distinct scattering regions along the line of sight[20,21].

As the survey frequently revisits the same positions, we can place strong constraints on burst repetition. We find no events at similar dispersion measures exceeding a signal-to-noise ratio of 9 at the dispersion measure and the positions of the ASKAP FRBs. Dwell times at these positions range from 8 days to 47 days (see Table 1), and there were 236–1,235 visits to the fields. In total, 12,456 hours of observations were conducted in the direction of the detected FRBs, and 61 days of observation of the FRB fields were conducted within ± 15 days of the FRB detection. For one of the detected bursts, FRB 171019, we conducted follow-up observations with the Parkes radio telescope two and three days after the FRB detection. In a total of 1 hour of observing, no pulses were detected at the dispersion measure of FRB 171019 above a limiting fluence of 1.5 Jy ms, which is a factor of 150 fainter than the FRB 171019 detection.

The fluence distribution of the ASKAP sample may be indicative of a cosmologically evolving population. We examined the distribution of the ASKAP sample with the V/V_{max} test statistic[22], which uses the measured signal-to-noise ratio (S/N) for each burst to assess the volume within which each burst was detected, relative to the maximum volume in which it could have been detected in Euclidean space-time. In such a population—distributed homogeneously over a volume small enough that the curvature of the Universe is negligible—the ensemble-averaged value of the statistic V/V_{max} would be 0.5. For the bursts in the

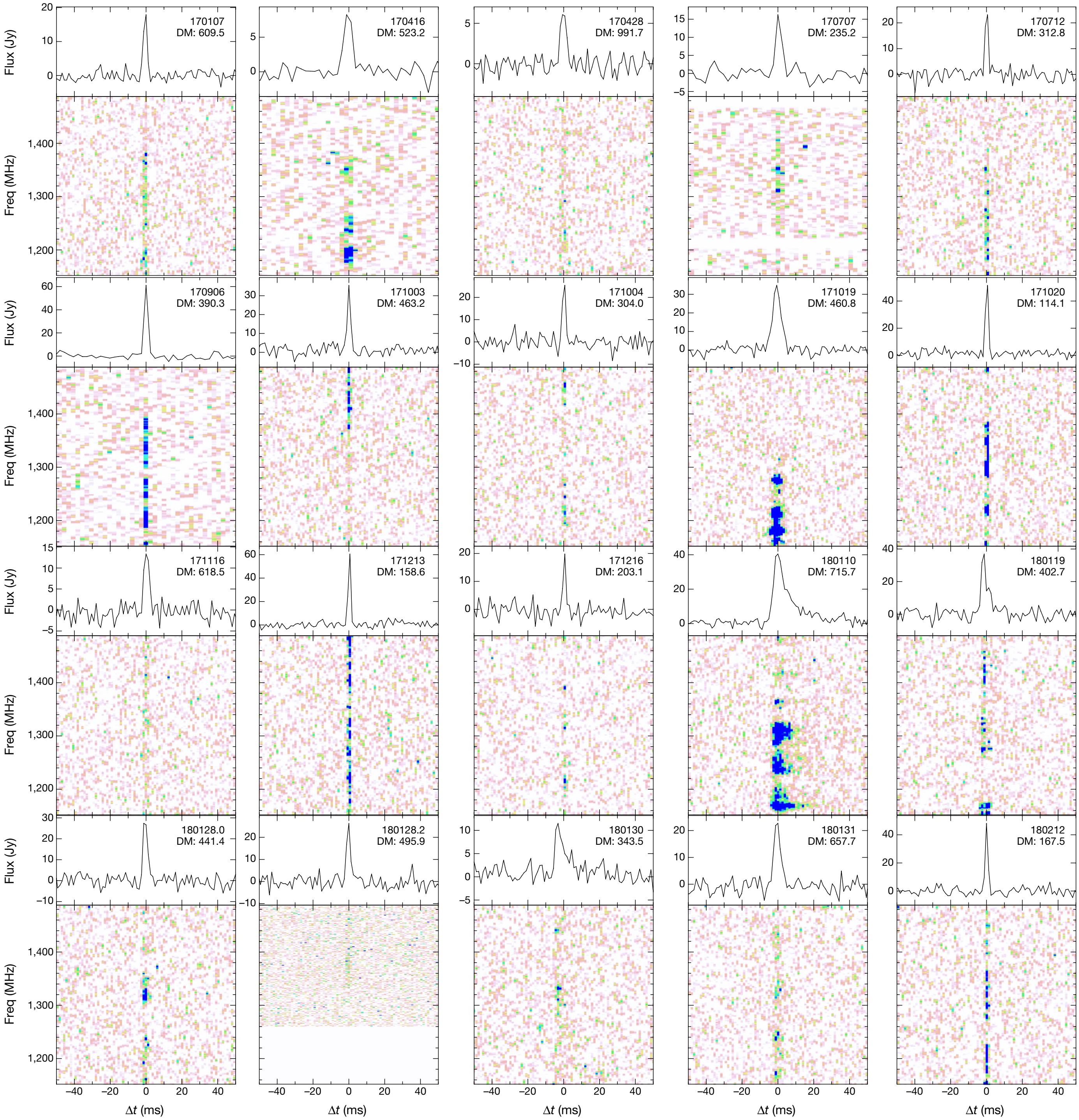


Fig. 1 | Pulse profiles and dynamic spectra of ASKAP FRBs. In the upper part of each panel the FRB name and dispersion measure (DM, in units of pc cm^{-3}) are shown, as well as the pulse profile. The lower part of each panel shows the FRB spectra, which have been dedispersed to the maximum-likelihood dispersion measure. The colour scale is set to range from the mean to 4σ of the off-pulse intensity.

ASKAP sample above our completeness threshold, we find $\langle V / V_{\max}\rangle = 0.58 \pm 0.07$. This is consistent with being produced by a Euclidean population with 12% confidence (see Supplementary Information, section 3). The one-sided probability was determined by simulations that model a realistic burst population with dispersion and widths consistent with the observed ASKAP population, detected by a system with our characteristics. For a pure power-law integral source count distribution, $N(>E) \propto E^{\alpha}$, parametrized by a spectral index α, the measured $V/V_{\max}$ value implies $\alpha = -2.1^{+0.6}_{-0.5}$ (67% confidence) over the range of fluences probed by ASKAP, evidence for steeper-than-Euclidean fluence distribution[23].

Comparison of the dispersion-measure distributions of the ASKAP and Parkes samples shows that dispersion measure is a distance indicator. The median dispersion measure of the ASKAP sample is 441 pc cm^{-3}, which is a factor of two smaller than that of the Parkes sample, 880 pc cm^{-3}. A Kolmogorov–Smirnov test finds that the probability that the two distributions are inconsistent is 99.9%. The difference in dispersion-measure distributions cannot be explained by the poorer time and frequency resolution of the ASKAP system (see Supplementary Information, section 5). This confirms both that there is both a correlation between dispersion measure and source fluence, and that dispersion measure can be used as a proxy for distance. However,

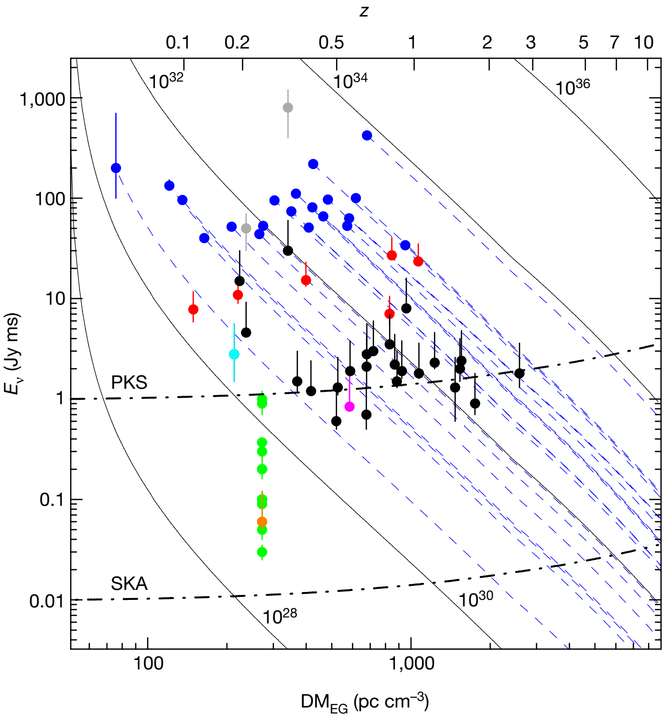


Fig. 2 | Distribution of FRB fluences and extragalactic dispersion measures. The extragalactic dispersion measures, DM_{EG}, are corrected for the inferred contribution of the Milky Way. The coloured circles denote FRBs detected with the ASKAP (blue), Parkes (black), UTMOST (red), Green Bank Telescope (magenta) and Arecibo (orange) radio telescopes. We also highlight the Parkes FRB candidate 010621 (cyan), which may be a Galactic source. Beam-corrected fluences have been estimated for two Parkes FRBs[18] (150807 and 010724) and are plotted in grey. Repeated pulses[29] from FRB 121102 are displayed in green. Uncertainties in beam fluence differ by telescope. For ASKAP and bursts from the repeating FRB 121102, we show 1σ (67% confidence) upper and lower limits. For other nonrepeating FRBs, lower limits are 1σ, but upper limits are twice the detected fluence, to reflect uncertainty in burst position within antenna pattern. The upper horizontal axis shows redshifts, assuming a homogenously distributed intergalactic plasma[10] and a host contribution of $50(1+z)^{-1}$ pc cm^{-3}, as discussed in the main text. The blue dashed curves show the fluences expected for the ASKAP-detected bursts if they were detected at larger distances (see Supplementary Information, section 4). The black curves show contours of constant spectral-energy density, in units of erg Hz^{-1}. The dash-dotted curves are lines of constant fluence, after accounting for redshift-dependent time dilation, denoting 10σ sensitivities of the Parkes radio telescope and the mid-frequency first-phase component of the future Square Kilometre Array[30] to 1-ms bursts. Further details of the data used can be found in Supplementary Information, section 4.

the difference in the distributions is smaller than would be expected for a non-evolving Euclidean population. For any individual burst, the average fluence decreases with the inverse square of distance. Because the Parkes sample is a factor of approximately 50 more sensitive than the ASKAP sample, it would be sensitive to the same source that is a factor of roughly 7 more distant. As we are comparing the median of the populations detected with Parkes and ASKAP, and not individual sources or standard candles, the ratio of dispersion measures is smaller—a consequence of a broad luminosity function for the population[24].

Figure 2 shows both the distribution of fluence plotted against extragalactic dispersion measures for all published FRBs, and the fluence/distance relationship—the latter assuming the model for host dispersion-measure contribution and extragalactic dispersion described above. The solid black curves are contours of constant energy density, calculated assuming pulses are isotropically beamed, and using the global spectral index to correct to the rest frame of the emitter. The dashed blue lines show the extrapolation of ASKAP FRB fluences to higher distances. Notably, the highest dispersion-measure event from Parkes[25]—FRB 160102—has an inferred energy comparable to those observed in ASKAP. On the basis of this extrapolation, the energies of the Parkes bursts overlap those of our sample, and are therefore more distant versions of the ASKAP events. The absence of sources above a spectral energy density of approximately 10^{34} erg Hz^{-1} for both the Parkes and the ASKAP samples is unlikely to be solely due to the frequency and temporal resolution of the data-recording systems, so could represent either a dwindling population or an energy cut-off (see Supplementary Information, section 5).

There are also marked differences between the ASKAP and Parkes burst populations, and the repeating FRB 121102. First, the ASKAP and Parkes samples show no evidence of repetition, despite large amounts of follow-up time and dense searches around the times of FRB detections. The repetition rate of FRB 121102 is intermittent, with frequent detections on month-long time scales, followed by similar length periods of apparent quiescence[3,26]. The absence of repetition enables us to reject, at the 99% confidence level, the hypothesis that all ASKAP FRBs repeat with the same properties[26] as FRB 121102 (see Supplementary Information, section 6). Second, the population of ASKAP bursts has a steep spectrum. While pulses from the repeating FRB show strong spectral modulation, equally energetic pulses are detected over a frequency range extending from 1.4 GHz to 8 GHz (ref. [27]). Furthermore, FRB 121102 is underluminous relative to the remaining bursts, as displayed in Fig. 2.

The results presented here build on previously noted differences between the repeating and the remaining non-repeating sources. For example, measurements of Faraday rotation suggest that luminous bursts propagate through dispersing plasma that is nonmagnetized[17], weakly magnetized[20,28], or has highly disordered magnetic fields. Such large Faraday rotations could still be hidden in the unpolarized FRBs, but it would be impossible to hide in the case of those with substantial polarization. By contrast, the repeating FRB source is found to have Faraday rotation (and hence magnetic-field strengths) more than four orders of magnitude larger[27]. We do not at present have the capability to measure Faraday rotation with ASKAP in FRB-search mode, but expect to upgrade these systems to make the necessary polarimetric observations shortly. We are also commissioning interferometric modes and expect to soon be able to localize detections to arcsecond accuracy. Unique identification of host galaxies will further distinguish between repeating and nonrepeating burst sources.

Code availability

The code used to conduct the FRB searches, FREDDA, will be publicly released shortly, but a pre-release version is available from K.W.B. (keith.bannister@csiro.au). Detections were processed using the dspsr (http://dspsr.sourceforge.net) and psrchive (http://psrchive.sourceforge.net) software packages for analysing pulsar and FRB data.

Data availability

Raw data files (totalling 1 PB) are archived on tape at the Pawsey Superconducting Centre. Cut-outs of the raw data, in pulsar filterbank format (http://sigproc.sourceforge.net), and posterior localization regions, are available on the CSIRO data access portal through https://doi.org/10.25919/5b6ae6b515850. Other data products are available on request from R.M.S.

Received: 8 April 2018; Accepted: 3 August 2018;
Published online 10 October 2018.

1. Lorimer, D. R., Bailes, M., McLaughlin, M. A., Narkevic, D. J. & Crawford, F. A bright millisecond radio burst of extragalactic origin. *Science* **318**, 777–780 (2007).
2. Petroff, E. et al. FRBCAT: the fast radio burst catalogue. *Publ. Astron. Soc. Aust.* **33**, e045 (2016).
3. Spitler, L. G. et al. A repeating fast radio burst. *Nature* **531**, 202–205 (2016).
4. Chatterjee, S. et al. A direct localization of a fast radio burst and its host. *Nature* **541**, 58–61 (2017).
5. Champion, D. J. et al. Five new fast radio bursts from the HTRU high-latitude survey at Parkes: first evidence for two-component bursts. *Mon. Not. R. Astron. Soc.* **460**, L30–L34 (2016).

6. McConnell, D. et al. The Australian Square Kilometre Array Pathfinder: performance of the Boolardy engineering test array. *Publ. Astron. Soc. Aust.* **33**, e042 (2016).
7. Bannister, K. W. et al. The cetection of an extremely bright fast radio burst in a phased array feed survey. *Astrophys. J.* **841**, L12 (2017).
8. Macquart, J.-P. & Johnston, S. On the paucity of fast radio bursts at low Galactic latitudes. *Mon. Not. R. Astron. Soc*. **451**, 3278–3286 (2015).
9. Xu, J. & Han, J. L. Extragalactic dispersion measures of fast radio bursts. *Res. Astron. Astrophys*. **15**, 1629 (2015).
10. Inoue, S. Probing the cosmic reionization history and local environment of gamma-ray bursts through radio dispersion. *Mon. Not. R. Astron. Soc*. **348**, 999–1008 (2004).
11. McQuinn, M. Locating the "missing" baryons with extragalactic dispersion measure estimates. *Astrophys. J.* **780**, L33 (2014).
12. Yao, J. M., Manchester, R. N. & Wang, N. A new electron-density model for estimation of pulsar and FRB distances. *Astrophys. J.* **835**, 29 (2017).
13. Meyer, M. J. et al. The HIPASS catalogue—I. Data presentation. *Mon. Not. R. Astron. Soc*. **350**, 1195–1209 (2004).
14. Jones, D. H. et al. The 6dF galaxy survey: final redshift release (DR3) and southern large-scale structures. *Mon. Not. R. Astron. Soc*. **399**, 683–698 (2009).
15. Cordes, J. M. et al. Lensing of fast radio bursts by plasma structures in host galaxies. *Astrophys. J.* **842**, 35 (2017).
16. Oslowski, S. et al. Real-time detection of an extremely high signal-to-noise ratio fast radio burst during observations of PSR J2124–3358. *Astron. Telegr.* 11385 (2018).
17. Ravi, V. et al. The magnetic field and turbulence of the cosmic web measured using a brilliant fast radio burst. *Science* **354**, 1249–1252 (2016).
18. Ravi, V. The observed properties of fast radio bursts. *Mon. Not. R. Astron. Soc*. (in the press); preprint at https://arxiv.org/abs/1710.08026 (2017).
19. Lambert, H. C. & Rickett, B. J. On the theory of pulse propagation and two-frequency field statistics in irregular interstellar plasmas. *Astrophys. J.* **517**, 299–317 (1999).
20. Masui, K. et al. Dense magnetized plasma associated with a fast radio burst. *Nature* **528**, 523–525 (2015).
21. Farah, W. et al. FRB microstructure revealed by the real-time detection of FRB170827. *Mon. Not. R. Astron. Soc*. (in the press); preprint at https://arxiv.org/abs/1803.05697 (2018).
22. Oppermann, N., Connor, L. D. & Pen, U.-L. The Euclidean distribution of fast radio bursts. *Mon. Not. R. Astron. Soc*. **461**, 984–987 (2016).
23. Macquart, J.-P. & Ekers, R. D. Fast radio burst event rate counts—I. Interpreting the observations. *Mon. Not. R. Astron. Soc*. **474**, 1900–1908 (2018).
24. von Hoerner, S. Radio source counts and cosmology. *Astrophys. J.* **186**, 741–766 (1973).
25. Bhandari, S. et al. The SUrvey for pulsars and extragalactic radio bursts—II. New FRB discoveries and their follow-up. *Mon. Not. R. Astron. Soc*. **475**, 1427–1446 (2018).
26. Law, C. J. et al. A multi-telescope campaign on FRB 121102: implications for the FRB Population. *Astrophys. J.* **850**, 76 (2017).
27. Michilli, D. et al. An extreme magneto-ionic environment associated with the fast radio burst source FRB 121102. *Nature* **553**, 182–185 (2018).
28. Keane, E. F. et al. The host galaxy of a fast radio burst. *Nature* **530**, 453–456 (2016).
29. Scholz, P. et al. The repeating fast radio burst FRB 121102: multi-wavelength observations and additional bursts. *Astrophys. J.* **833**, 177 (2016).
30. Macquart, J. P. et al. Fast transients at cosmological distances with the SKA. *Adv. Astrophys. Square Kilometre Array (AASKA14) Proc. Sci*. **215**, 55 (2015).

Acknowledgements We thank the Australia Telescope National Facility (ATNF) engineering and technical staff for their help in supporting these observations, and especially thank the staff of the Murchison Radio-astronomy observatory. We thank C. Flynn, P. Edwards, N. Tejos and V. McIntyre for comments on the manuscript, and members of the Commensal Real-time ASKAP Fast Transients (CRAFT) team for discussions. We thank the Murchison Widefield Array (MWA) principal engineer, R. Wayth, for access to the Galaxy supercomputer graphics processing units (GPU) cluster. R.M.S. and S.O. acknowledge Australian Research Council (ARC) grant FL150100148. R.M.S. also acknowledges support through ARC grant CE170100004. G.G. acknowledges support through a Commonwealth Scientific and Industrial Research Organisation (CSIRO) Office of the Chief Executive (OCE) postdoctoral fellowship. Parts of this research were conducted by the ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO; grant CE110001020). This research was also supported by the ARC through grant DP18010085. The Australian SKA Pathfinder and Parkes radio telescopes are part of the ATNF, which is managed by the CSIRO. Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Centre. Establishment of ASKAP, the Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund. We acknowledge the Wajarri Yamatji people as the traditional owners of the Observatory site. This research has made use of the National Aeronautics and Space Administration (NASA)/Infrared Processing and Analysis Center (IPAC) Extragalactic Database (NED), which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA.

Reviewer information *Nature* thanks J. Cordes, D. Lorimer and S. Ransom for their contribution to the peer review of this work.

Author contributions K.W.B. led the development of the CRAFT data-acquisition system. R.M.S., J.-P.M. and K.W.B. designed the survey. R.M.S., J-.P.M., K.W.B. and R.D.E drafted the manuscript. R.M.S. and K.W.B. conducted the observations, with assistance from A.W.H. and M.A.V. K.W.B. designed the search code. K.W.B., C.W.J., S.O., H.Q. and M.S. verified survey efficiency. R.M.S., with discussions with J.R.A., implemented the FRB localization algorithm. R.M.S., J.-P.M. and R.D.E interpreted the fluence and dispersion-measure distributions of the population. C.W.J., S.O. and J.-P.M. interpreted the nonrepetition of the ASKAP sample and compared it with the repeating FRB. R.M.S. and S.O. led searches for follow-up bursts at Parkes. E.M.S. studied the optical fields surrounding the detected FRBs. R.J.B., M.B., A.J.B., J.D.B., A.P.C., C.H., A.W.H., M.L., M.M., D.M., M.A.P, E.R.T., J.T., M.A.V. and M.T.W. contributed to development and commissioning of the CRAFT observing mode. J.R.A., C.S.A., M.E.B., J.D.C., G.G., G.H. and C.J.R. contributed to ASKAP commissioning and early science.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0588-y.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.M.S. or J.-P.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

https://doi.org/10.1038/s41586-018-0605-1

Space-borne Bose–Einstein condensation for precision interferometry

Dennis Becker[1,16], Maike D. Lachmann[1,16], Stephan T. Seidel[1,15,16], Holger Ahlers[1], Aline N. Dinkelaker[2], Jens Grosse[3,4], Ortwin Hellmig[5], Hauke Müntinga[3], Vladimir Schkolnik[2], Thijs Wendrich[1], André Wenzlawski[6], Benjamin Weps[7], Robin Corgier[1,8], Tobias Franz[7], Naceur Gaaloul[1], Waldemar Herr[1], Daniel Lüdtke[7], Manuel Popp[1], Sirine Amri[8], Hannes Duncker[5], Maik Erbe[9], Anja Kohfeldt[9], André Kubelka-Lange[3], Claus Braxmaier[3,4], Eric Charron[8], Wolfgang Ertmer[1], Markus Krutzik[2], Claus Lämmerzahl[3], Achim Peters[2], Wolfgang P. Schleich[10,11,12,13], Klaus Sengstock[5], Reinhold Walser[14], Andreas Wicht[9], Patrick Windpassinger[6] & Ernst M. Rasel[1]*

Owing to the low-gravity conditions in space, space-borne laboratories enable experiments with extended free-fall times. Because Bose–Einstein condensates have an extremely low expansion energy, space-borne atom interferometers based on Bose–Einstein condensation have the potential to have much greater sensitivity to inertial forces than do similar ground-based interferometers. On 23 January 2017, as part of the sounding-rocket mission MAIUS-1, we created Bose–Einstein condensates in space and conducted 110 experiments central to matter-wave interferometry, including laser cooling and trapping of atoms in the presence of the large accelerations experienced during launch. Here we report on experiments conducted during the six minutes of in-space flight in which we studied the phase transition from a thermal ensemble to a Bose–Einstein condensate and the collective dynamics of the resulting condensate. Our results provide insights into conducting cold-atom experiments in space, such as precision interferometry, and pave the way to miniaturizing cold-atom and photon-based quantum information concepts for satellite-based implementation. In addition, space-borne Bose–Einstein condensation opens up the possibility of quantum gas experiments in low-gravity conditions[1,2].

Studies of quantum systems such as matter-waves in the presence of a gravitational field[3] can help to improve our understanding of general relativity[4] and quantum mechanics. Because the sensitivity of measuring inertial forces with matter-wave interferometers is proportional to the square of the time that the atoms spend in the interferometer[5], an extended free-fall of atoms in the interferometer results in a large enhancement in sensitivity[1,6]. In this context, slowly spreading ensembles with pico- or femtokelvin-scale expansion energies, obtained by Bose–Einstein condensation[7,8] in combination with 'delta-kick' collimation[9–11], remain in the interferometer for longer and are therefore essential for interferometry over timescales of the order of tens of seconds. The associated large coherence lengths of the ensemble are needed to combine precision with accuracy[1].

Generating and manipulating Bose–Einstein condensates (BECs) with low expansion energies is difficult because they can easily be compromised by experimental imperfections, very small forces and gravity. By creating a BEC on board a sounding rocket, we successfully demonstrated key atom-optics methods under microgravity conditions. Our experimental apparatus[12–14] (Fig. 1) is equipped with a multilayer atom chip[15–17]. Its in-flight performance compares well with ground-based demonstrations, producing about 10^5 atoms in 1.6 s. This flux made it possible to perform a large number of experiments during the space flight, exemplified here by images of a space-based BEC (Fig. 1e) and of Bragg scattering of a BEC (Fig. 1f). The latter shows the spatial density profile of the BEC and its replica, which was generated by Bragg scattering at a light crystal and moves with a relative velocity that corresponds to the transfer of two photon recoils. In Fig. 1f we compare the size of the BEC in terms of the Thomas–Fermi radius and its separation from its replica 70 ms after the Bragg scattering event, which occurred 15.6 ms after the release of the BEC from the atom chip. The expansion velocity of the BEC is nine times smaller than the velocity that is transferred during Bragg scattering. The stripe pattern results from an intensity modulation of the light fields that induce the Bragg scattering.

In Fig. 2 we summarize the experiments of the MAIUS-1 mission that were performed in space and during the launch of the rocket. These experiments build on those of the QUANTUS collaboration[18,19], and complement those on dual-species interferometry[20] and those that involve clocks based on laser-cooled atoms[21]. They are also instrumental for NASA's Cold Atom Laboratory[2] (CAL) on the International Space Station (ISS) and for the NASA-DLR Bose–Einstein Condensate and Cold Atom Laboratory (BECCAL) multi-user facility, which is currently in the planning phase[22].

Here we report on BEC experiments with rubidium-87 atoms in space. We studied the phase transition from a thermal ensemble to a BEC by adjusting the temperature via forced radio-frequency evaporation of thermal atoms out of the atom-chip magnetic trap. In Fig. 3a we show the spatial atomic density of the thermal ensemble and the BEC at three different final radio frequencies of the forced evaporation (at the final cooling step). During the phase transition, with decreasing temperature the number of atoms in the thermal ensemble (extracted using a Gaussian fit, red curve in Fig. 3a) decreases markedly whereas that in the BEC increases (parabolic fit, blue curve in Fig. 3a). In Fig. 3b, c we compare the formation of BECs in space and on the ground; we also plot the fraction of atoms in the BEC with respect to the total atom number.

The comparison reveals that, for the same final radio frequency, the observed ratio of thermal and condensed atoms (and hence the fraction of the total number of atoms in the BEC) was lower in space than on the ground. We suspect that this difference is due to a change in the magnetic field in space with respect to that on the ground, resulting from, for example, a thermal drift in the current supply. In addition, the numbers of atoms in the thermal ensemble and in the BEC in space

[1]Institute of Quantum Optics, QUEST-Leibniz Research School, Leibniz University Hannover, Hanover, Germany. [2]Department of Physics, Humboldt-Universität zu Berlin, Berlin, Germany. [3]Center of Applied Space Technology and Microgravity (ZARM), University of Bremen, Bremen, Germany. [4]Institute of Space Systems, German Aerospace Center (DLR), Bremen, Germany. [5]Institute of Laser-Physics, University Hamburg, Hamburg, Germany. [6]Institute of Physics, Johannes Gutenberg University Mainz (JGU), Mainz, Germany. [7]Simulation and Software Technology, German Aerospace Center (DLR), Brunswick, Germany. [8]Institut des Sciences Moléculaires d'Orsay (ISMO), CNRS, Université Paris-Sud, Université Paris-Saclay, Orsay, France. [9]Ferdinand-Braun-Institut, Leibniz-Institut für Höchstfrequenztechnik, Berlin, Germany. [10]Institut für Quantenphysik and Center for Integrated Quantum Science and Technology (IQST), Ulm, Germany. [11]Hagler Institute for Advanced Study, Texas A&M University, College Station, TX, USA. [12]Texas A&M AgriLife Research, Texas A&M University, College Station, TX, USA. [13]Institute for Quantum Science and Engineering (IQSE), Department of Physics and Astronomy, Texas A&M University, College Station, TX, USA. [14]Institut für Angewandte Physik, Technische Universität Darmstadt, Darmstadt, Germany. [15]Present address: OHB System AG, Weßling, Germany. [16]These authors contributed equally: Dennis Becker, Maike D. Lachmann, Stephan T. Seidel. *e-mail: rasel@iqo.uni-hannover.de

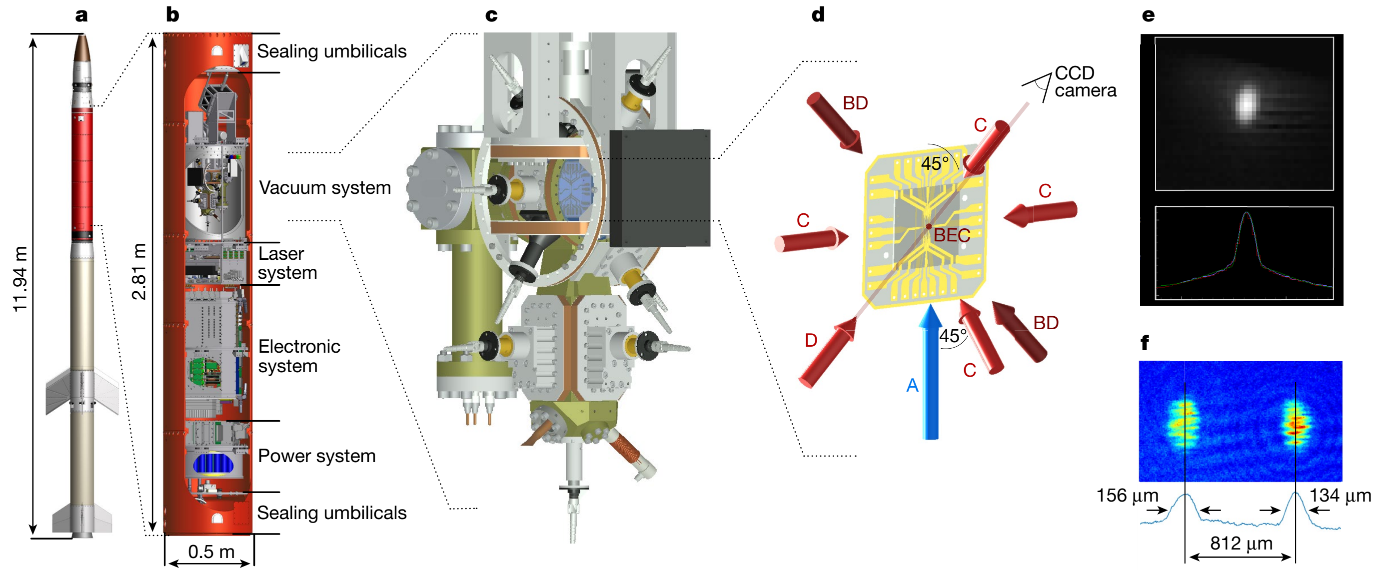


Fig. 1 | Set-up for space-borne Bose–Einstein condensation. **a**–**d**, The rocket (**a**) carried the payload (**b**), including the vacuum system (**c**) that houses the atom chip (**d**), into space. On the atom chip, a magneto-optical trap formed by laser beams (C) is first loaded from the cold atomic beam (A). Afterwards, the BEC is created in, transported by and released from the magnetic trap of the atom chip. Two additional light beams (BD) induce Bragg diffraction, and a charge-coupled device (CCD) camera records the absorption image of the BEC using laser light (D). **e**, Grey-scale absorption image of the spatial density of the BEC in space (top; white corresponds to the highest densities) and its one-dimensional density profile (bottom; integrated from the top to the bottom of the image), which were sent to ground control in low resolution. **f**, Our demonstration of Bragg scattering, apparent in the momentum distribution of the BEC, opens up a path towards atom interferometry in space. The image contrasts the size of the BECs in the spatial superposition that we created with their relative separation 70 ms after the transfer of two photon recoils onto the replica, which moves to the right. The colour scale shows the spatial density of the clouds (blue, low; red, high).

are 64% higher than those obtained on the ground. This improvement in the BEC flux is most probably due to more efficient loading into the magnetic trap in the absence of gravitational sag. To optimize the BEC flux even further, the circuitry of the multilayer atom chip offers various trap configurations, with variable volume and depth. However,

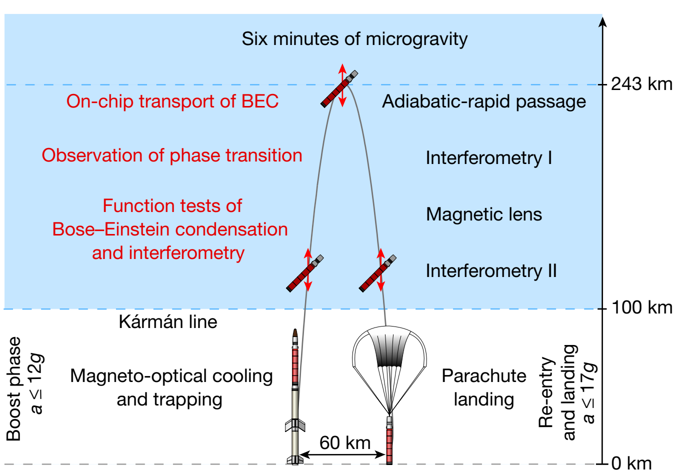


Fig. 2 | Schedule for the MAIUS-1 sounding-rocket mission. During the boost phase (bottom left) and the 6 min of space flight (blue-shaded region), 110 atom-optics experiments were performed. Those discussed here are printed in red. In space (above the Kármán line, 100 km above the ground), inertial perturbations are reduced to a few parts per million of gravity, the pointing of the length axis is stabilized with respect to gravity (indicated by the red arrows) and the spin of the rocket is suppressed to about 5 mrad s^{-1} owing to rate control. During re-entry, the peak forces on the payload (a) exceed the gravitational force on the ground (g) by a factor of up to 17.

experiments of this kind require more time than was available during our flight.

Because transporting and shaping BECs to create compact wavepackets are key to interferometry, we investigated the evolution of the BEC in free fall after release and the transport of BECs on the atom chip away from the surface of the chip via its impact on the BEC motion in free fall. In space, and therefore in the absence of gravitational sag, we can compare the predictions of a theoretical simulation directly with the observations. The BECs were moved across a distance of 0.8 mm from the surface of the chip. For this purpose, the homogeneous magnetic field, which in combination with the atom chip determines the location of the Ioffe–Pritchard trap, was lowered smoothly over 50 ms with a sigmoidal time dependence.

In particular, we studied oscillations in the centre-of-mass position of a BEC excited by its transport on the atom chip. For this purpose, the BEC was kept trapped for variable hold times of up to 25 ms before it was released. In Fig. 4a we show the positions with respect to the surface of the chip of BECs detected 50 ms after release as a function of hold time; these positions reflect the varying initial velocities of the BECs due to their centre-of-mass oscillation. Using these data, we can reconstruct the motion of the BEC in the trap. Of a total of ten measurements, five tested the repeatability of the preparation for zero hold time (Fig. 4a, green circles) and five probed the oscillatory behaviour (black circles) for increasing hold time; the latter illustrate the sinusoidal dependence of the distance of the centre-of-mass of the BEC from the chip on hold time, over various trials, consistent with the fitted sinusoidal behaviour (dashed purple line) of a trapped quantum gas.

In addition, we investigated the motion of the BEC for times of up to 300 ms after release after zero hold time. We include the data from Fig. 4a for 50 ms after release (green circles) also in Fig. 4b. According to Fig. 4a, the velocity of the BEC associated with the oscillation is maximum for zero hold time. This velocity adds to that due to transport away from the chip; hence, the total velocity of the BEC after release

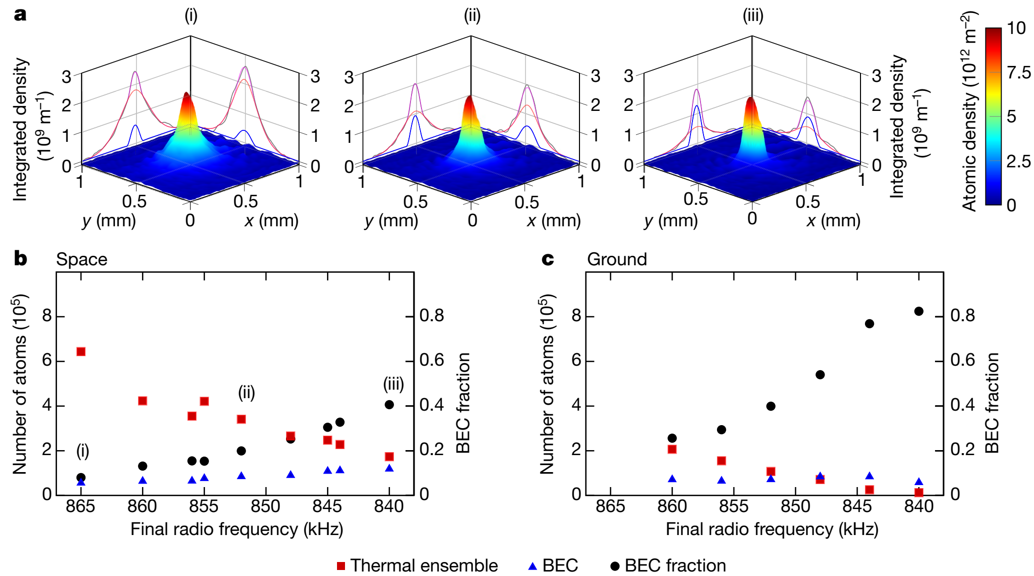


Fig. 3 | Phase transition to the BEC in space and on the ground, controlled by the final radio frequency of the forced evaporation. **a**, Spatial atomic density (colour scale) and corresponding line integrals (solid grey lines), as well as Gaussian (red lines) and parabolic (blue lines) fits of the line integrals of the thermal and condensed atoms, respectively, and their sum (violet lines), for cases in space where 8% (i), 20% (ii) and 41% (iii) of the atoms are in the BEC state. **b**, **c**, The number of magnetically trapped atoms in the thermal ensemble (red squares, left axis) is higher in space (**b**) than on the ground (**c**), resulting in more atoms in the BEC (blue triangles, left axis) in space; for a comparable BEC fraction, there are 64% more atoms in the BEC in space than in the BEC on the ground. The dependence of the fraction of the total number of atoms in the BEC (that is, the number of atoms in the BEC divided by the sum of the numbers of atoms in the BEC and in the thermal ensemble; black circles, right axis) on the radio frequency is also different in space and on the ground. In **b**, cases (i)–(iii) from **a** are indicated for reference.

was as large as 8.8 mm s^{-1}, as inferred from a linear fit (dashed purple line in Fig. 4b). Moreover, after release, we transferred the atoms by adiabatic rapid passage into a mixture of the Zeeman states so that we could detect possible residual fields. Despite the strong effect of preparation and transport on the motion of the BEC, the trajectories demonstrate only a small scatter in the experimental data for different Zeeman states of the $F=2$ manifold (Fig. 4b; triangles, $m_F=0$; circles, $m_F=2$; m_F is the magnetic angular momentum quantum number; F is the hyperfine splitting quantum number) and set an upper bound of 1% for the corresponding relative fluctuation in the initial velocity that is possibly caused by the various manipulations of the atom chip. In addition to minimizing the amplitude of the centre-of-mass oscillations, phase stability of these oscillations is required for quantum tests such as those proposed for the STE-QUEST satellite mission[1].

We compare our measurements to a theoretical model[23] of the dynamics of the BEC after creation, including its oscillation in the trap, release and evolution until detection. This model includes the current-carrying wire structures of the experimental set-up and solves the Gross–Pitaevskii equation in the Thomas–Fermi regime[24]. Our experimental results shown in Fig. 4a, b agree well with our simulation (solid blue lines), differing only by a slight underestimate of the oscillation amplitude and a corresponding small velocity offset. This difference may result from the model for shutting off the release trap as we have limited knowledge of the magnetic field dynamics during the switch-off. Most other potential reasons are excluded by the simulation, which allows us to check for the influence on the BEC dynamics of the circuitry of the atom chip, of the Helmholtz coils and of uncertainties in the current values (shaded areas in Fig. 4).

The motion on the atom chip causes complicated oscillations in the shape of the BEC[25] (Fig. 4c); this finding again demonstrates the importance of phase-stable manipulations. Our theoretical simulations show that the observed variations in the size of the BEC as defined by the Thomas–Fermi radii originate from oscillations in the shape of the BEC rather than from experimental noise, as confirmed by the low phase scatter at release (Fig. 4d), with relative fluctuations of only about 2%. Although the density of the BEC reduces by one order of magnitude during transport, the remaining mean-field energy still causes the BEC to expand by up to about 1 mm in diameter after 300 ms (Fig. 4d), in agreement with our theoretical prediction (solid blue and red lines). The thermal background of the released atoms has a temperature of approximately 100 nK and the residual expansion of the BEC corresponds to a kinetic energy equivalent to a few nanokelvin, as expected from the decompression during transport. Further reduction of the expansion energies requires additional measures, such as delta-kick collimation, which could not be studied during the short rocket flight.

The phase stability of the centre-of-mass motion and of the collective size oscillations of the BEC is necessary to reproducibly release the atoms into free fall with vanishing velocity and to lower the expansion to the required level. To minimize the release velocity, we propose transport protocols that largely suppress the amplitude of the sinusoidal oscillations, bringing into reach BECs with release velocities below micrometres per second. Low expansion rates in all three dimensions, corresponding to kinetic energies of a few tens of picokelvin, are possible with atom chips by combining delta-kick collimation with excitation of a quadrupole shape oscillation[23]. The latter compensates for the asymmetric trap of the atom chip favouring the implementation of cylindrical lenses.

In conclusion, we used a payload on board a sounding rocket to create BECs in space. Although many experiments were performed during this space flight, here we focus on studies of the phase transition of the thermal ensemble to a BEC and the collective oscillations of the centre-of-mass and size of the BEC in the trap induced by the transport of the BEC away from the atom chip. The reproducibility that we have demonstrated allows the implementation of more sophisticated transport protocols (such as shortcut-to-adiabaticity protocols[23] or optimal control protocols), and enables shape oscillations to be used jointly with delta-kick collimation to reduce and shape the expansion of BECs, to extend the time that they spend in the interferometer. Our experiments demonstrate the atom-optics tools that are required for satellite gravimetry[26], for quantum tests of the equivalence principle[27] and for gravitational-wave detection based on matter-wave interferometry in space[28]. Moreover, they pave the way to miniaturizing cold-atom and photon-based quantum information concepts, and to integrating these concepts into quantum-communication satellites[29–33].

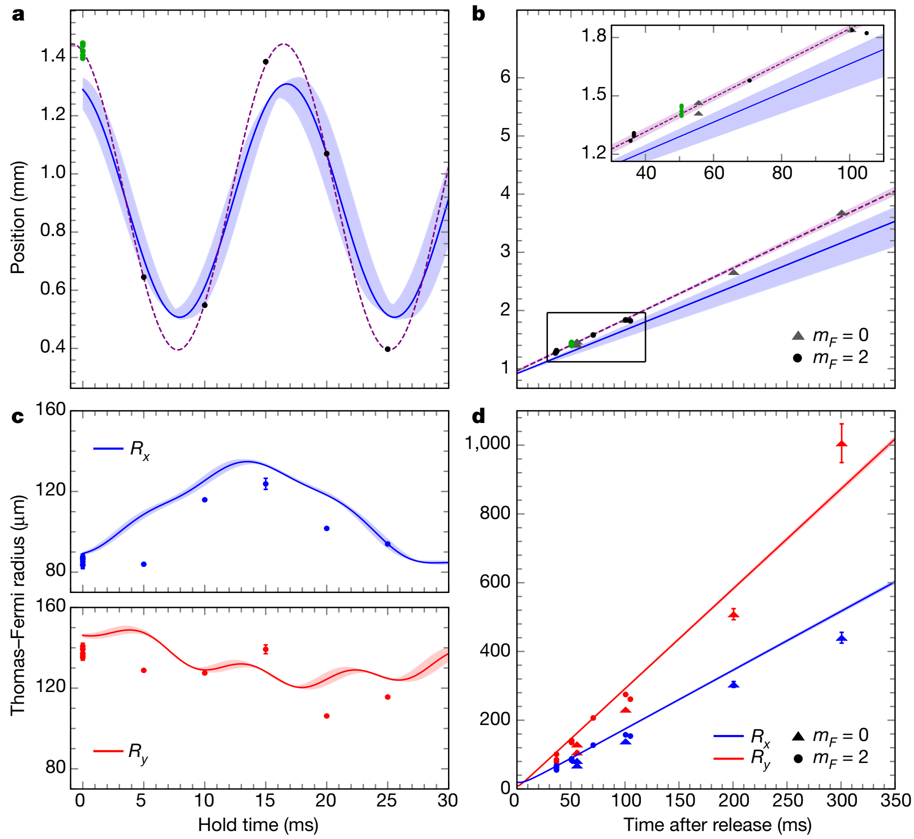


Fig. 4 | Excitation of the centre-of-mass motion and oscillations in the shape of a space-borne BEC as a result of its transport away from an atom chip. **a**, From the modulation of the distance travelled by the BEC 50 ms after its release for different hold times, we infer the centre-of-mass motion of the BEC in the trap as a function of hold time by fitting a sinusoid (purple dashed line) to the data (green and black circles for immediate release and varying hold times, respectively). The simulation of the evolution of the BEC (blue line) agrees well with the data, but underestimates the amplitude of the oscillation. **b**, The centre-of-mass motion of the BEC away from the atom chip after release from the trap is well fitted by a linear function (purple dashed line; purple shading indicates the 95% confidence interval), and is almost identical for different Zeeman states of the $F=2$ manifold (grey triangles, $m_F=0$; black and green circles, $m_F=2$; green circles in **a** and **b** represent the same data). The simulation of the dynamics of the BEC based on the Gross–Pitaevskii equation in the Thomas–Fermi limit is also shown (blue line). The inset shows a close-up of the boxed region of the main plot. **c**, The Thomas–Fermi radii R_x (top, blue circles) and R_y (bottom, red circles) serve as measures of the size and thus the shape of the BEC 50 ms after release. For varying hold time, these radii display complicated oscillations, which also appear in our simulations (red and blue lines) of the BEC evolution. **d**, Thomas–Fermi radii for condensates that were released immediately after transport and freely expanded. After 300 ms, the BEC has grown in size, up to about 1 mm. Most experiments were performed with BECs in the $m_F=0$ state (blue and red triangles), with the results in accordance with our theory for BECs in the $m_F=0$ state (red and blue lines), but some were performed with BECs in the $m_F=2$ state (blue and red circles). Possible deviations due to residual magnetic field gradients are below the measurement resolution. In all panels, error bars indicate uncertainties related to fitting the images of the BECs. Uncertainties in the theoretical model (blue and red shaded areas) reflect the degree of knowledge of the experimental parameters, such as those related to the generation of a magnetic field by electrical circuits and currents (in particular, the range of currents used in the simulations; see Methods).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0605-1.

Received: 18 June 2018; Accepted: 9 August 2018;
Published online 17 October 2018.

1. Aguilera, D. N. et al. STE-QUEST—test of the universality of free fall using cold atom interferometry. *Class. Quantum Gravity* **31**, 115010 (2014).
2. Elliott, E. R., Krutzik, M. C., Williams, J. R., Thompson, R. J. & Aveline, D. C. NASA's Cold Atom Lab (CAL): system development and ground test status. *npj Microgravity* **4**, 16 (2018).
3. Colella, R., Overhauser, A. W. & Werner, S. A. Observation of gravitationally induced quantum interference. *Phys. Rev. Lett.* **34**, 1472–1474 (1975).
4. Misner, C. W., Thorne, K. S. & Wheeler, J. A. *Gravitation* (Princeton Univ. Press, Princeton, 1973).
5. Berman, P. R. *Atom Interferometry* Ch. 9 (Academic Press, New York, 1997).
6. Dimopoulos, S., Graham, P. W., Hogan, J. M. & Kasevich, M. A. Testing general relativity with atom interferometry. *Phys. Rev. Lett.* **98**, 111102 (2007).
7. Cornell, E. A. & Wieman, C. E. Nobel lecture: Bose-Einstein condensation in a dilute gas, the first 70 years and some recent experiments. *Rev. Mod. Phys.* **74**, 875–893 (2002).
8. Ketterle, W. Nobel lecture: When atoms behave as waves: Bose-Einstein condensation and the atom laser. *Rev. Mod. Phys.* **74**, 1131–1151 (2002).
9. Chu, S., Bjorkholm, J. E., Ashkin, A., Gordon, J. P. & Hollberg, L. W. Proposal for optically cooling atoms to temperatures of the order of 10^{-6} K. *Opt. Lett.* **11**, 73–75 (1986).
10. Müntinga, H. et al. Interferometry with Bose-Einstein condensates in microgravity. *Phys. Rev. Lett.* **110**, 093602 (2013).
11. Kovachy, T. et al. Matter wave lensing to picokelvin temperatures. *Phys. Rev. Lett.* **114**, 143004 (2015).
12. Grosse, J. et al. Design and qualification of an UHV system for operation on sounding rockets. *J. Vac. Sci. Technol. A* **34**, 031606 (2016).
13. Schkolnik, V. et al. A compact and robust diode laser system for atom interferometry on a sounding rocket. *Appl. Phys. B* **122**, 217 (2016).
14. Kubelka-Lange, A. et al. A three-layer magnetic shielding for the MAIUS-1 mission on a sounding rocket. *Rev. Sci. Instrum.* **87**, 063101 (2016).
15. Hänsel, W., Hommelhoff, P., Hänsch, T. W. & Reichel, J. Bose-Einstein condensation on a microelectronic chip. *Nature* **413**, 498–501 (2001).
16. Folman, R., Krüger, P., Schmiedmayer, J., Denschlag, J. & Henkel, C. Microscopic atom optics: from wires to an atom chip. *Adv. At. Mol. Opt. Phys.* **48**, 263–356 (2002).
17. Fortágh, J. & Zimmermann, C. Magnetic microtraps for ultracold atoms. *Rev. Mod. Phys.* **79**, 235–289 (2007).
18. van Zoest, T. et al. Bose-Einstein condensation in microgravity. *Science* **328**, 1540–1543 (2010).
19. Rudolph, J. et al. A high-flux BEC source for mobile atom interferometers. *New J. Phys.* **17**, 065001 (2015).

20. Barrett, B. et al. Dual matter-wave inertial sensors in weightlessness. *Nat. Commun*. **7**, 13786 (2016).
21. Liu, L. et al. In-orbit operation of an atomic clock based on laser-cooled ^{87}Rb atoms. *Nat. Commun*. **9**, 2760 (2018).
22. Gibney, E. Universe's coolest lab set to open up quantum world. *Nature* **557**, 151–152 (2018).
23. Corgier, R. et al. Fast manipulation of Bose-Einstein condensates with an atom chip. *New J. Phys*. **20**, 055002 (2018).
24. Pethick, C. & Smith, H. *Bose-Einstein Condensation in Dilute Gases* Ch. 6 (Cambridge Univ. Press, Cambridge, 2002).
25. Stringari, S. Collective excitations of a trapped Bose-condensed gas. *Phys. Rev. Lett*. **77**, 2360–2363 (1996).
26. Douch, K., Wu, H., Schubert, C., Müller, J. & dos Santos, F. P. Simulation-based evaluation of a cold atom interferometry gradiometer concept for gravity field recovery. *Adv. Space Res*. **61**, 1307–1323 (2018).
27. Altschul, B. et al. Quantum tests of the Einstein equivalence principle with the STE-QUEST space mission. *Adv. Space Res*. **55**, 501–524 (2015).
28. Hogan, J. M. et al. An atomic gravitational wave interferometric sensor in low earth orbit (AGIS-LEO). *Gen. Relativ. Gravit*. **43**, 1953–2009 (2011).
29. Armengol, J. M. P. et al. Quantum communications at ESA: towards a space experiment on the ISS. *Acta Astronaut*. **63**, 165–178 (2008).
30. Ren, J.-G. et al. Ground-to-satellite quantum teleportation. *Nature* **549**, 70–73 (2017).
31. Liao, S.-K. et al. Satellite-to-ground quantum key distribution. *Nature* **549**, 43–47 (2017).
32. Yin, J. et al. Satellite-based entanglement distribution over 1200 kilometers. *Science* **356**, 1140–1144 (2017).
33. Yin, J. et al. Satellite-to-ground entanglement-based quantum key distribution. *Phys. Rev. Lett*. **119**, 200501 (2017).

Acknowledgements This work is supported by the DLR Space Administration with funds provided by the Federal Ministry for Economic Affairs and Energy (BMWi) under grant numbers DLR 50WM1131-1137, 50WM0940 and 50WM1240. W.P.S. thanks Texas A&M University for a Faculty Fellowship at the Hagler Institute for Advanced Study at Texas A&M University and Texas A&M AgriLife for support for this work. The research of the IQST is financed partially by the Ministry of Science, Research and Arts Baden-Württemberg. N.G. acknowledges funding from Niedersächsisches Vorab through the Quantum- and Nano-Metrology (QUANOMET) initiative within the project QT3. W.H. acknowledges funding from Niedersächsisches Vorab through the project Foundations of Physics and Metrology project. R.C. is a recipient of DAAD (Procope action and mobility scholarship) and a member of the IP@ Leibniz programme, which is supported by LU Hanover. S.T.S. is grateful for non-monetary support from DLR MORABA before, during and after the MAIUS-1 launch. We thank E. Kajari and M. Eckardt for the chip model code and A. Roura and W. Zeller for their input. We thank C. Spindeldreier and H. Blume from IMS Hanover for FPGA software development. We acknowledge the contributions of PTB Brunswick and LNQE Hanover towards fabricating the atom chip. We thank ESRANGE Kiruna and DLR MORABA Oberpfaffenhofen for assistance during the test and launch campaign.

Reviewer information *Nature* thanks L. Liu and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.B., M.D.L., S.T.S., H.A., A.N.D., J.G., O.H., H.M., V.S., T.W., A.We., B.W., T.F., D.L., M.P., M.E., A.K., H.D., A.K.-L. and M.K. designed, built and tested the apparatus. D.B., M.D.L., H.A., A.N.D., J.G., O.H., H.M., V.S., T.W., A.We. and B.W., with S.T.S. as scientific lead, planned and executed the campaign. D.B., M.D.L. and S.T.S. evaluated the data. N.G., R.C., E.C., S.A., W.H. and D.B. carried out the simulations. E.M.R., W.P.S., M.D.L., D.B. and N.G. wrote the manuscript, with contributions from all authors. C.B., W.E., C.L., A.P., W.P.S., K.S., R.W., A.Wi. and P.W. are the co-principal investigators of the project, and E.M.R. its principal investigator.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0605-1.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to E.M.R.

METHODS

Detection and analysis of absorption images. BECs were detected using an absorption-detection system that uses the $F=2 \rightarrow F'=3$ transition of rubidium-87 atoms. A collimated detection beam enters the vacuum chamber via a window from one side, excites the atoms, and leaves the chamber at the opposite side. With two lenses and one mirror, the beam is directed onto the sensor of a CCD camera (pco.1400). During each sequence, one image including the shadow of the atoms and a later one without the shadow is recorded. Both are corrected with a dark image in the absence of the detection beam. The atomic column density is calculated from the optical density in MATLAB according to ref. [34]. Using a bimodal fit, consisting of a Thomas–Fermi and a Gaussian distribution, the number of atoms in the ensemble and its size and position are calculated.

Simulation. The theory curves in Fig. 4 are generated using a Gross–Pitaevskii equation solver, in the Thomas–Fermi regime, as described elsewhere[3]. The trapping frequencies are extracted from a detailed atom-chip simulation package, as used previously[19]. The chip structure is comparable to that obtained in ref. [19] and involves two z currents on two layers of the chip (base chip and science chip) and two currents through coils. Possible lack of knowledge of these currents is the origin of the shaded areas in Fig. 4. We scan these currents (from larger to smaller values) over the following intervals: x coil, $[-0.02, -0.01]$ A; y coil, $[-1.605, -1.595]$ A; base chip, $[4.95, 5]$ A; science chip, $[1.98, 2]$ A.

Data availability

Source Data for Figs. 3b, c and 4 are available with the online version of the paper.

34. Reinaudi, G., Lahaye, T., Wang, Z. & Guéry-Odelin, D. Strong saturation absorption imaging of dense clouds of ultracold atoms. *Opt. Lett.* **32**, 3143–3145 (2007).

https://doi.org/10.1038/s41586-018-0544-x

Subcycle observation of lightwave-driven Dirac currents in a topological surface band

J. Reimann[1], S. Schlauderer[2], C. P. Schmid[2], F. Langer[2], S. Baierl[2], K. A. Kokh[3,4], O. E. Tereshchenko[4,5], A. Kimura[6], C. Lange[2], J. Güdde[1], U. Höfer[1]* & R. Huber[2]*

Harnessing the carrier wave of light as an alternating-current bias may enable electronics at optical clock rates[1]. Lightwave-driven currents have been assumed to be essential for high-harmonic generation in solids[2–6], charge transport in nanostructures[7,8], attosecond-streaking experiments[9–16] and atomic-resolution ultrafast microscopy[17,18]. However, in conventional semiconductors and dielectrics, the finite effective mass and ultrafast scattering of electrons limit their ballistic excursion and velocity. The Dirac-like, quasi-relativistic band structure of topological insulators[19–29] may allow these constraints to be lifted and may thus open a new era of lightwave electronics. To understand the associated, complex motion of electrons, comprehensive experimental access to carrier-wave-driven currents is crucial. Here we report angle-resolved photoemission spectroscopy with subcycle time resolution that enables us to observe directly how the carrier wave of a terahertz light pulse accelerates Dirac fermions in the band structure of the topological surface state of Bi_2Te_3. While terahertz streaking of photoemitted electrons traces the electromagnetic field at the surface, the acceleration of Dirac states leads to a strong redistribution of electrons in momentum space. The inertia-free surface currents are protected by spin–momentum locking and reach peak densities as large as two amps per centimetre, with ballistic mean free paths of several hundreds of nanometres, opening up a realistic parameter space for all-coherent lightwave-driven electronic devices. Furthermore, our subcycle-resolution analysis of the band structure may greatly improve our understanding of electron dynamics and strong-field interaction in solids.

The band structure of a crystalline solid describes the nontrivial energy–momentum relation of electrons that results from the overlap of orbitals of adjacent atoms. Because the band structure governs key physical and chemical properties of the solid, a complete understanding of it is important both for fundamental materials science and for applications. For instance, energy gaps distinguish insulators from metals, whereas the band slope sets the electron velocity. Angle-resolved photoemission spectroscopy (ARPES) can be used to map out the occupied part of band structures by measuring the kinetic energy of photoemitted electrons as a function of their momenta. One of the successes of ARPES is the experimental observation of three-dimensional topological

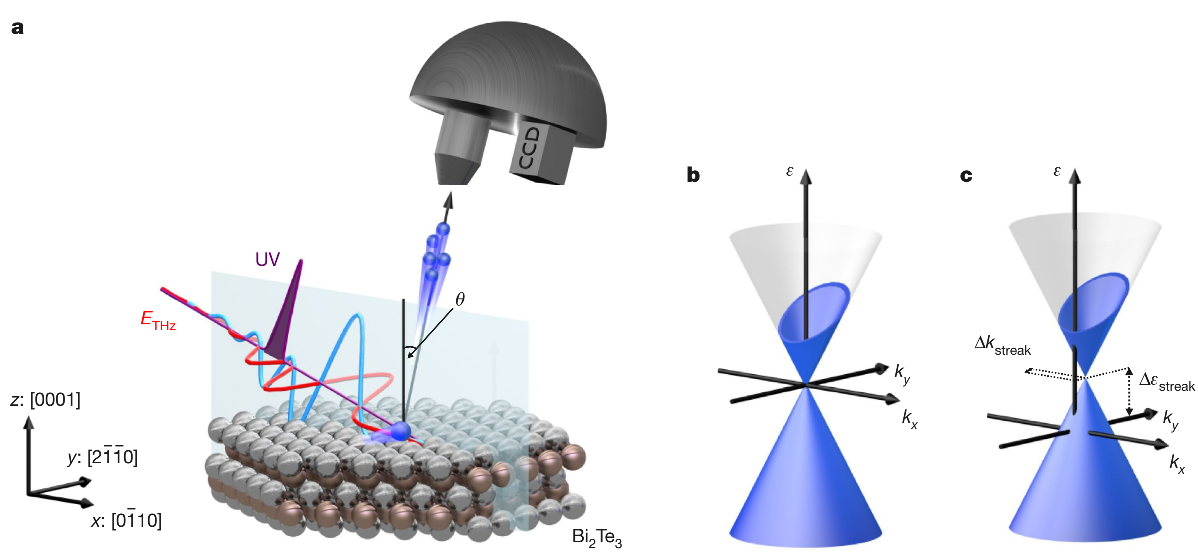


Fig. 1 | Concept of subcycle time-resolved ARPES. a, Electrons (blue spheres) in the topological surface state of Bi_2Te_3 (crystal lattice; brown spheres, Bi atoms; grey spheres, Te atoms) are accelerated by an intense linearly polarized THz field E_{THz} (red waveform, s-polarization; blue waveform, p-polarization) and are photoemitted by an ultrashort time-delayed p-polarized UV pulse (violet). The kinetic energy ε_{kin} and emission angle θ of the liberated electrons are measured by a hemispherical electron analyser to determine the binding energy and the parallel momentum k_y of the electrons in the topological surface band along the $\overline{\Gamma}-\overline{K}$ direction. The plane of incidence (x–z plane) is indicated by a transparent rectangle. **b**, Acceleration of Dirac fermions in the topological surface state can shift the Fermi surface. **c**, The interaction of the photoemitted electrons with the THz field in vacuum can manifest as a streaking of the photoemission spectra with an energy shift $\Delta\varepsilon_{streak}$ and/or a momentum shift Δk_{streak} of the apparent band structure.

[1]Department of Physics, Philipps-University of Marburg, Marburg, Germany. [2]Department of Physics, University of Regensburg, Regensburg, Germany. [3]V.S. Sobolev Institute of Geology and Mineralogy, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. [4]Novosibirsk State University, Novosibirsk, Russia. [5]A.V. Rzhanov Institute of Semiconductor Physics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. [6]Graduate School of Science, Hiroshima University, Hiroshima, Japan. *e-mail: hoefer@physik.uni-marburg.de; rupert.huber@ur.de

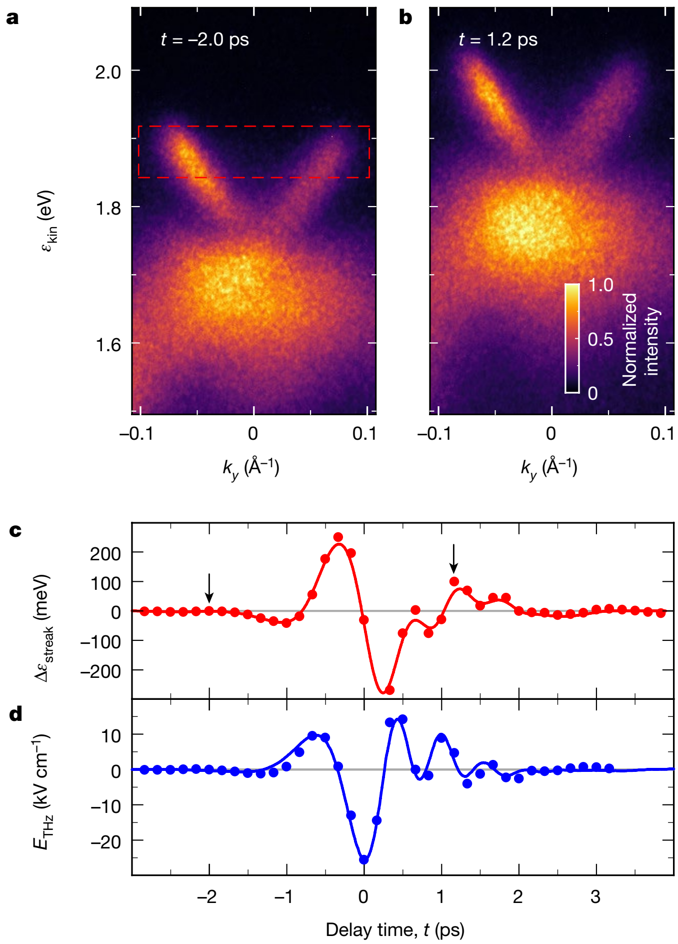


Fig. 2 | Energy streaking of the photoelectrons by p-polarized THz radiation. a, Photoemission map taken before the arrival of the THz field ($t = -2.0$ ps). The dashed red rectangle depicts the region that is traced to evaluate the time-dependent energy shift. **b**, Corresponding photoemission map taken 1.2 ps after the arrival of the maximum THz electric field, exhibiting THz streaking. **c**, Energy shift of the photoemission spectra as a function of the pump–probe delay time (red circles, experiment; solid curve, simulation; see text). The arrows indicate the delay times at which the photoemission maps in **a** and **b** are taken. **d**, Reconstructed electric field amplitude of the standing wave in front of the surface (blue circles). The blue solid curve shows the scaled and time-shifted electric field as measured externally by electro-optic sampling.

insulators[19–21] with unusual transport properties: owing to strong spin–orbit interaction, these solids are insulating in the bulk, but the surface exhibits gapless states that are protected by time-reversal symmetry[19]. The quasi-relativistic dispersion and a reduction in scattering, caused by spin–momentum locking, makes these surface states promising for use in ultrafast low-loss electronics. Yet, the motion of Dirac fermions driven by an electric field is not accessible by conventional photoemission spectroscopy.

By exploiting ultrashort laser pulses, time-resolved ARPES has been used to trace the dynamics of photo-injected currents at surfaces as a small imbalance of carriers moving in opposite directions[30]. ARPES has also revealed light-induced Floquet states[23] and interband transitions[22], specifically in topological insulators. Furthermore, attosecond metrology has been used to access subcycle time delays in the momentum-integrated photoemission from surfaces with known band structures[9–13,16]. A systematic understanding of electromagnetically driven intraband currents in materials with unknown electronic states and dynamics, such as topological insulators, requires an analysis of the band structure with subcycle resolution.

Here, we present a time-resolved ARPES study with subcycle resolution. We observe directly how an intense terahertz (THz) wave accelerates Dirac fermions in the quasi-relativistic band structure of the topological surface state of Bi_2Te_3. The inertia-free transient Dirac currents peak at a density of 2 A cm^{-1} with a maximum ballistic excursion of electrons in real space of several hundred nanometres. Our results open up a new chapter in lightwave electronics, whereby currents induced by optical carrier waves can be traced directly within the band structure.

The concept of our experiment is illustrated in Fig. 1a. A single crystal of Bi_2Te_3 is kept in an ultrahigh-vacuum (UHV) chamber. This material belongs to a class of three-dimensional topological insulators that exhibit a topological surface state with a single non-degenerate Dirac cone at the $\overline{\Gamma}$ point in the Brillouin zone[19,20]. A strong THz electric field (E_{THz}) is focused onto the surface to accelerate the Dirac fermions. This process is expected to shift the electron occupation in momentum space along the field direction (Fig. 1b). These dynamics should be strongly influenced by many-body interactions, such as scattering. In our experiment, we probe the transient electron distribution directly in momentum space via time- and angle-resolved photoemission. We use a time-delayed ultraviolet (UV) laser pulse (duration, 100 fs; centre wavelength, 201 nm) to release electrons out of the surface. An electrostatic hemispherical electron analyser, equipped with a charge-coupled device (CCD) detector, then images the relationship between the energy ε and the momentum k_y of the photoelectrons parallel to the surface (Fig. 1b). On their way to the detector, the electrons also interact with the THz field in the vacuum, so the apparent band structure is effectively offset in energy and momentum (Fig. 1c). Because the duration of the UV pulse is much shorter than the oscillation period of the THz driving field, this 'streaking' effect can be used to sample the electric field close to the surface, in a similar way to the sampling of near-infrared lightwaves by attosecond pulses[9–13,16]. The relative strength of intraband acceleration and streaking in vacuum can be controlled by the THz polarization.

In the first step, we show that THz fields polarized in the plane of incidence (p-polarization) predominantly lead to energy streaking of photoemitted electrons. In Fig. 2a, b we display angle-resolved photoemission spectra for two different delay times t between the p-polarized THz field and the UV probe. The characteristic V-shaped dispersion of Dirac fermions demonstrates the presence of a topological surface state, whereas the bulk valence band manifests as the broad dispersion at lower energies. The Fermi velocity that we deduce from the linear dispersion of the topological surface state, $\varepsilon(k_y) = \hbar v_F k_y$, is $v_F = 4.1$ Å fs^{-1} = 410 nm ps^{-1}. The THz field appears to offset the entire band structure in energy, while leaving its shape and momentum position largely unchanged. This behaviour is well explained considering that, for our Bi_2Te_3 sample, the Fresnel reflection at the surface (see Methods) leads to strong suppression of the THz field component in the sample plane and strong enhancement of the perpendicular component. The resulting, dominantly out-of-plane field causes pronounced energy streaking of photoemitted electrons, but cannot effectively accelerate Dirac fermions within the surface bands.

For a quantitative analysis, we extract the THz field from the experimental energy streaking $\Delta\varepsilon_{streak}(t)$ (Fig. 2c, red circles), which we determine by tracing the energy position of a fixed cut-out of the photoemission intensity distribution (Fig. 2a, red dashed rectangle) as a function of time. For a given THz transient, $\Delta\varepsilon_{streak}(t)$ can be calculated by integrating the classical equation of motion of an electron accelerated by the Lorentz force (see Methods). By inverting this relation, we retrieve the THz waveform $E_{THz}(t)$ above the sample surface (Fig. 2d, blue circles). An integration of the equation of motion in the full three-dimensional THz field including the strongly screened in-plane component (Fig. 2c, red line) is in excellent agreement with the experimental shift $\Delta\varepsilon_{streak}(t)$, confirming that the dynamics is dominated by the out-of-plane electric field. We also verify that the reconstructed THz transient coincides with the waveform determined

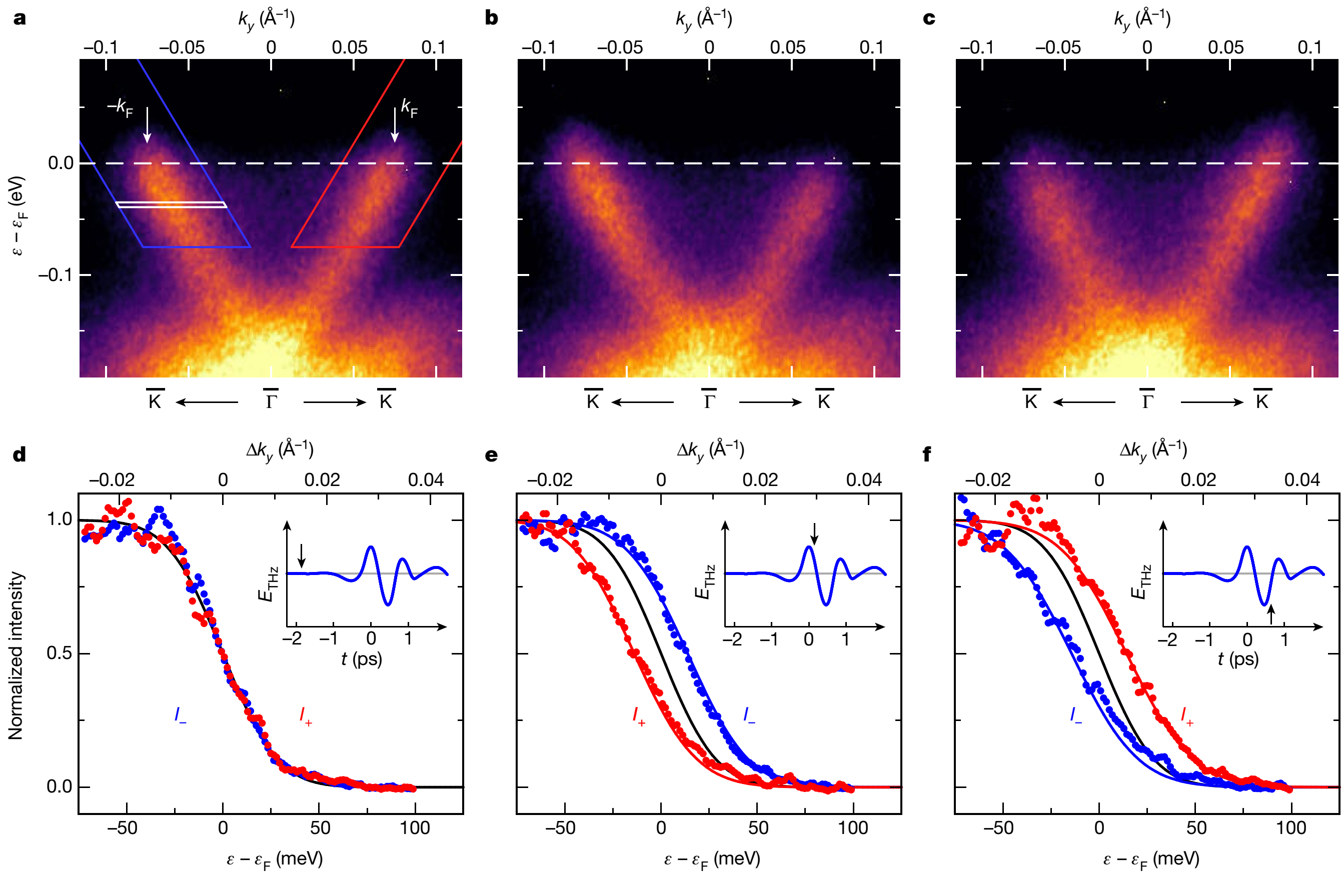


Fig. 3 | Acceleration of the electrons within the surface band by s-polarized THz radiation. a–c, Photoemission maps before the arrival of the THz field ($t = -1.86$ ps) (**a**), just after the first positive maximum of the electric field ($t = 0.14$ ps) (**b**) and immediately after the negative crest of the electric field ($t = 0.64$ ps) (**c**). The white dashed lines indicate the Fermi level for the unperturbed system; the Fermi wave vector k_F is indicated in **a**. **d–f**, Intensity distributions I_- and I_+ of the photoemission maps shown in **a–c** along the left (blue circles) and right (red circles) branches of the Dirac cone, respectively. I_- and I_+ are obtained by integrating the measured intensity over horizontal slices (see, for example, white box in **a**) moved along trapezoidal regions as depicted in **a**. The solid curves show the results of simulations based on the Boltzmann equation. The black curves depict the equilibrium situation and correspond to a Fermi–Dirac distribution at 80 K convoluted with the experimental energy resolution. The insets in **d–f** show the reconstructed electric field, with the black arrows indicating the delay times of the corresponding snapshots.

by electro-optic sampling outside the UHV chamber (Fig. 2d, blue line), which demonstrates the reliability of the field retrieval by streaking.

The electric field of s-polarized THz pulses is oriented along the y direction, in which our electron analyser detects the electron momentum k_y. Although streaking is strongly suppressed in this geometry because of screening of the THz field parallel to the surface, the remaining momentum streaking Δk_{streak} still allows us to probe the instantaneous field directly (see Methods). Moreover, the s-polarized, in-plane component may accelerate electrons within the topologically protected surface band. In Fig. 3a–c we display snapshots of photoemission spectra for three characteristic delay times t: before the THz transient (Fig. 3a), 0.14 ps after the positive field maximum (Fig. 3b) and 0.18 ps after the negative crest (Fig. 3c). For vanishing fields (Fig. 3a), the left and right branches of the Dirac cone are occupied up to the same Fermi energy ($\varepsilon_F \approx 200$ meV above the Dirac point) and Fermi momentum ($k_F = 0.075$ Å^{-1}). When the electric field is applied, the occupation becomes asymmetric.

More quantitatively, we extract the intensity distributions of photoelectrons along the left (I_-; blue circles) and right (I_+; red circles) branches of the Dirac cone (Fig. 3d–f). For each energy, $I_\pm$ is obtained by integrating the photoemission intensity over a narrow momentum interval (such as depicted by the white box in Fig. 3a) centred about the Dirac band. The normalized distributions $I_\pm(t)$ directly represent femtosecond snapshots of the population in each branch of the topological surface state, that is, the energy and momentum distributions $f(\varepsilon, t)$ and $f(k_y, t)$, respectively. Whereas the red and blue curves coincide with each other in the absence of an electric field (Fig. 3d), they are shifted horizontally in energy and momentum with respect to each other when the electric field is present (Fig. 3e, f). This shift is reversed when the electric field changes sign. The opposite shift of the population of the surface band for opposite electron momenta signifies an ultrafast displacement of the Fermi circle in momentum space. To the best of our knowledge, this marks the first direct observation of electron currents driven by the carrier wave of an ultrashort electromagnetic pulse in the band structure of a solid, in general, and of Dirac currents, in particular. As shown in Supplementary Video 1, we can even retrieve a complete quasi-continuous subcycle video of THz-driven Dirac fermions within their band structure. The transient shift of the electrons in momentum space results in a net current flowing in the positive or negative y direction on the surface of the sample.

The measured momentum distribution $f(k_y, t)$ allows us to identify the key mechanisms that govern the subcycle transport of the Dirac fermions directly in the time domain. Before the interaction with the THz field (Fig. 3d), the measured $I_\pm(k_y)$ is well described by a Fermi–Dirac distribution for our sample temperature of 80 K (solid line). Shortly after the field maxima (Fig. 3e, f), the curves are broadened and strongly displaced from their equilibrium distribution. Here, the transient electric field $E_{THz}(t)$ coherently accelerates the electrons out of equilibrium, while scattering limits their ballistic motion. This scenario can be described approximately by the semi-classical Boltzmann equation (see Methods), as shown by the blue and red solid lines in Fig. 3e, f.

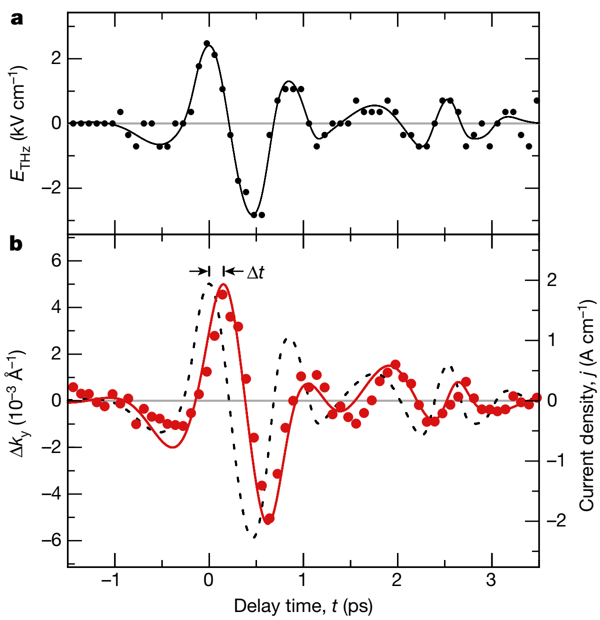


Fig. 4 | Dynamics of electric current within the surface band. **a**, Electric field parallel to the surface as reconstructed from the transient momentum shift of the photoelectrons (black circles). The black solid line is a spline through the data. **b**, Temporal evolution of the current density extracted from the photoemission intensity distributions (red circles). The red solid curve shows the simulated current dynamics calculated for scattering times of $\tau_R = \tau_{k\bar{k}} = 1$ ps and the electric field transient of **a**, which is depicted in **b** by the dashed curve, for comparison.

Our model accounts for electron scattering in a simple relaxation-time approximation including Pauli exclusion. The scattering time for processes that relax the excess energy of an electron is τ_R; the time for elastic scattering between $+k_y$ and $-k_y$ is quantified by $\tau_{k\bar{k}}$. In the presence of a driving field, both effects tend to push the carriers towards a Fermi–Dirac distribution that is shifted in energy and momentum by the driving field (see Methods); the influence of scattering is remarkably weak in our experiment. This becomes obvious when we fit our model calculation to reproduce the experiment by adapting τ_R and $\tau_{k\bar{k}}$ and by scaling the electric field strength within the estimated experimental uncertainty. We achieve quantitative agreement (see, for example, Fig. 3e) only if both τ_R and $\tau_{k\bar{k}}$ are kept above 1 ps. Assuming scattering times as long as 2 ps (ref. [22]) deteriorates the agreement with the experiment only slightly, whereas shorter scattering times result in substantially different carrier distributions and/or weaker deflection than observed experimentally (see Methods).

The low scattering rate allows for a large displacement of the Fermi circle even for moderate peak fields of 2.4 kV cm^{-1} (Fig. 3b). In combination with the quasi-relativistic band structure, this situation results in extremely large current densities. With a maximum observed momentum shift of 0.005 Å^{-1}, we retrieve a peak surface current density of up to 2 A cm^{-1} (see Methods). This current flows within an atomically thin surface sheet that hosts the wavefunction of the topological surface state. Assuming a thickness of 1 nm, the corresponding bulk current density reaches values of up to 2×10^7 A cm^{-2}, or 4 nA per atom of the surface layer.

By repeating this analysis for various delay times t, we map out the subcycle dynamics of the current density $j(t)$ (Fig. 4b, red circles) together with the instantaneous THz field (Fig. 4b, dashed curve; Fig. 4a, black circles). The current density clearly shows the fingerprints of ballistic acceleration: $j(t)$ increases monotonically during the onset of the first intense THz half cycle. Although $E_{THz}(t)$ decreases for $t > 0$ ps, $j(t)$ keeps rising and reaches its maximum at $t = 0.15$ ps, shortly before the reversal of the driving field. This behaviour, which also occurs for all other THz half cycles, is expected only when the influence of scattering on the Dirac current is weak. Solving the Boltzmann equation for the experimentally measured THz driving field reproduces the current dynamics very well, including the delayed response and the details of the subsequent current oscillations (Fig. 4b, solid curve). Again, we find the optimum agreement for scattering times τ_R and $\tau_{k\bar{k}}$ above 1 ps. The actual scattering times may well be even longer, but the limited time window of a THz half cycle makes determining an upper bound challenging. These values readily exceed coherence times observed previously for lightwave-driven electron dynamics in semiconductors and dielectrics[2–6] by two orders of magnitude and attest to a reduced scattering phase space due to spin–momentum locking. It would be particularly interesting to contrast these results with future subcycle ARPES of Dirac fermions in graphene, which lacks spin–momentum locking.

These unique transport dynamics of topological Dirac fermions elevate carrier-wave electronics to the all-coherent level. In contrast to the massive quasiparticles that populate the parabolic bands of conventional dielectrics, the acceleration of quasi-relativistic topological surface states is inertia-free because the group velocity of Dirac fermions is very high from the outset. Hence, THz-accelerated Dirac fermions may propagate coherently over several hundred nanometres before undergoing scattering (see Extended Data Fig. 5). This scale exceeds electron excursions in dielectrics[2–6] and the gate width of state-of-the-art electronic transistors by orders of magnitude. Therefore, we anticipate that devices based on three-dimensional topological insulators will soon exploit all-coherent electron transfer at THz or even petahertz clock rates, and that the underlying subcycle currents may become accessible to complementary real-space measurements[27].

In conclusion, we resolve the momentum distribution of Dirac fermions directly as they are accelerated by the carrier wave of a THz pulse within the band structure of the topological surface state. The combination of quasi-relativistic dispersion and spin–momentum locking makes topological insulators ideal for ultrafast electronics: inertia-free charge currents and unprecedented coherence lengths move microelectronic devices driven by lightwaves in a broad spectral range into practical reach. The extremely low dissipation rate of the currents that we observed practically eliminates limitations imposed by Joule heating (see Methods), so the speed of future devices is limited only by the optical clock rate of the light pulse, which could be scaled to ever higher frequencies, facilitating ultimately fast electronics. The linear dispersion relation of Dirac fermions may even enable dispersion-free wave-packet motion. Owing to spin–momentum locking, the ballistic Dirac currents should carry a spin current, which should enable spintronics up to optical clock rates. Finally, our concept of subcycle ARPES provides a way of observing carrier transport directly in non-trivial band structures. This idea may herald a new era of time-domain investigations of surface and bulk band structures of new materials and phenomena, ranging from topology to high-temperature superconductivity.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0544-x.

Received: 20 March 2018; Accepted: 26 July 2018;
Published online 26 September 2018.

1. Krausz, F. & Stockman, M. I. Attosecond metrology: from electron capture to future signal processing. *Nat. Photon.* **8**, 205–213 (2014).
2. Vampa, G. et al. Linking high-harmonics from gases and solids. *Nature* **522**, 462–464 (2015).
3. Hohenleutner, M. et al. Real-time observation of interfering crystal electrons in high-harmonic generation. *Nature* **523**, 572–575 (2015).
4. Langer, F. et al. Lightwave-driven quasiparticle collisions on a subcycle timescale. *Nature* **533**, 225–229 (2016).
5. Garg, M. et al. Multi-petahertz electronic metrology. *Nature* **538**, 359–363 (2016).
6. Liu, H. et al. High-harmonic generation from an atomically thin semiconductor. *Nat. Phys.* **13**, 262–265 (2017).
7. Rybka, T. et al. Sub-cycle optical phase control of nanotunnelling in the single-electron regime. *Nat. Photon.* **10**, 667–670 (2016).

8. Higuchi, T., Heide, C., Ullmann, K., Weber, H. B. & Hommelhoff, P. Light-field-driven currents in graphene. *Nature* **550**, 224–228 (2017).
9. Miaja-Avila, L. et al. Laser-assisted photoelectric effect from surfaces. *Phys. Rev. Lett.* **97**, 113604 (2006).
10. Cavalieri, A. L. et al. Attosecond spectroscopy in condensed matter. *Nature* **449**, 1029–1032 (2007).
11. Schultze, M. et al. Attosecond band-gap dynamics in silicon. *Science* **346**, 1348–1352 (2014).
12. Neppl, S. et al. Direct observation of electron propagation and dielectric screening on the atomic length scale. *Nature* **517**, 342–346 (2015).
13. Locher, R. et al. Energy-dependent photoemission delays from noble metal surfaces by attosecond interferometry. *Optica* **2**, 405–410 (2015).
14. Siek, F. et al. Angular momentum-induced delays in solid-state photoemission enhanced by intra-atomic interactions. *Science* **357**, 1274–1277 (2017).
15. Feist, A. et al. Quantum coherent optical phase modulation in an ultrafast transmission electron microscope. *Nature* **521**, 200–203 (2015).
16. Tao, Z. et al. Direct time-domain observation of attosecond final-state lifetimes in photoemission from solids. *Science* **353**, 62–67 (2016).
17. Cocker, T. L. et al. An ultrafast terahertz scanning tunnelling microscope. *Nat. Photon.* **7**, 620–625 (2013).
18. Cocker, T. L., Peller, D., Yu, P., Repp, J. & Huber, R. Tracking the ultrafast motion of a single molecule by femtosecond orbital imaging. *Nature* **539**, 263–267 (2016).
19. Hasan, M. Z. & Kane, C. L. Colloquium: Topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
20. Zhang, H. J. et al. Topological insulators in Bi_2Se_3, Bi_2Te_3 and Sb_2Te_3 with a single Dirac cone on the surface. *Nat. Phys.* **5**, 438–442 (2009).
21. Chen, Y. L. et al. Experimental realization of a three-dimensional topological insulator, Bi_2Te_3. *Science* **325**, 178–181 (2009).
22. Kuroda, K., Reimann, J., Güdde, J. & Höfer, U. Generation of transient photocurrents in the topological surface state of Sb_2Te_3 by direct optical excitation with midinfrared pulses. *Phys. Rev. Lett.* **116**, 076801 (2016).
23. Mahmood, F. et al. Selective scattering between Floquet–Bloch and Volkov states in a topological insulator. *Nat. Phys.* **12**, 306–310 (2016).
24. Olbrich, P. et al. Room-temperature high-frequency transport of Dirac fermions in epitaxially grown Sb_2Te_3- and Bi_2Te_3-based topological insulators. *Phys. Rev. Lett.* **113**, 096601 (2014).
25. Souma, S. et al. Direct measurement of the out-of-plane spin texture in the Dirac-cone surface state of a topological insulator. *Phys. Rev. Lett.* **106**, 216803 (2011).
26. Sobota, J. A. et al. Ultrafast optical excitation of a persistent surface-state population in the topological insulator Bi_2Se_3. *Phys. Rev. Lett.* **108**, 117403 (2012).
27. Kastl, C., Karnetzky, C., Karl, H. & Holleitner, A. W. Ultrafast helicity control of surface currents in topological insulators with near-unity fidelity. *Nat. Commun.* **6**, 6617 (2015).
28. Minami, Y. et al. Terahertz-induced acceleration of massive Dirac electrons in semimetal bismuth. *Sci. Rep.* **5**, 15870 (2015).
29. Braun, L. et al. Ultrafast photocurrents at the surface of the three-dimensional topological insulator Bi_2Se_3. *Nat. Commun.* **7**, 13259 (2016).
30. Güdde, J., Rohleder, M., Meier, T., Koch, S. W. & Höfer, U. Time-resolved investigation of coherently controlled electric currents at a metal surface. *Science* **318**, 1287–1291 (2007).

Acknowledgements We thank R. Höfer for discussions. The work in Marburg was supported by the Deutsche Forschungsgemeinschaft (DFG) through SFB 1083 and grant number HO 2295/7 (SPP 1666). Work in Regensburg was supported by the DFG through grant numbers HU 1598/2-1 and SFB 1277 (Project A05) and by the European Research Council through grant number 305003 (QUANTUMsubCYCLE). O.E.T. and K.A.K. were supported by the Russian Science Foundation (project number17-12-01047). A.K. was financially supported by KAKENHI number 17H06138.

Reviewer information *Nature* thanks I. Katayama, S. Zhou and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.G., U.H. and R.H. conceived the study. J.R., S.S., C.P.S., F.L., S.B. and J.G. carried out the experiment. K.A.K., O.E.T. and A.K. provided the samples and performed the transport measurements. J.R., C.L., J.G. and U.H. carried out the theoretical modelling. All authors analysed the data, discussed the results and contributed to writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0544-x.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0544-x.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to U.H. or R.H.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Experimental set-up. We start with near-infrared pulses (centre wavelength, 807 nm; pulse energy, 5.5 mJ; pulse duration, 33 fs) from a titanium:sapphire amplifier (repetition rate, 3 kHz). Part of the laser output is used to generate intense THz pulses (pulse energy, 1 μJ) by tilted-pulse front optical rectification[31] in a cryogenically cooled lithium niobate ($LiNbO_3$) crystal (Extended Data Fig. 1a). A pair of wire-grid polarizers controls the polarization direction of the THz field. A typical transient after transmission through the fused silica window of the vacuum chamber is shown in Extended Data Fig. 1b. A second branch of the laser output is frequency-converted into the UV to release photoelectrons from the sample surface. To this end, we generate the fourth harmonic by cascaded sum-frequency generation with the fundamental amplifier pulses in β-barium borate (β-BBO) crystals (Extended Data Fig. 1c). Extended Data Fig. 1d depicts a typical resulting spectrum. A prism sequence separates the fourth harmonic from the fundamental pulses and pre-compensates for the material dispersion of the window of the vacuum chamber. The THz and UV pulses are spatially overlapped with a custom-built gold mirror featuring an aperture for the transmission of the UV pulses. A mechanical delay stage in the excitation beam path allows us to temporally delay THz and UV pulses with respect to each other. The fluence of the UV light is set by detuning the wave plate in front of the third-harmonic generation crystal (see Extended Data Fig. 1c). The collinear THz-pump and p-polarized UV-probe pulses are transmitted through a UV-grade fused silica window into a μ–metal shielded UHV chamber at a base pressure of 7×10^{-11} mbar. A spherical UV-grade aluminium mirror (focal length, 75 mm) mounted within the UHV chamber is used to focus both beams under an angle of incidence of 53° onto the sample. Photoelectrons are collected along the high-symmetry line $\overline{\Gamma}-\overline{K}$ perpendicular to the plane of light incidence by a hemispherical electron analyser with a display-type detector (Specs Phoibos 150). The energy resolution of the set-up is 45 meV (full-width at half-maximum), mainly limited by the bandwidth of the UV pulses. Energy and momentum shifts of photoelectrons can be determined with an accuracy of 3 meV and 0.004 Å^{-1}, respectively.

Sample preparation and characterization. The Bi_2Te_3 samples are fabricated by using the modified Bridgman method[32]. By choosing an appropriate solidification condition, a gradation in the carrier concentration was realized in an ingot naturally forming a p–n junction. Transport measurements within the temperature range of 10–300 K reveal a carrier mobility of about 10^4 cm^2 V^{-1} s^{-1} at 4 K, indicating the high crystal quality. The longitudinal resistance at 80 K is about 0.4 mΩ cm^{-1}. It increases with temperature, confirming bulk conductive characteristics. A clean and well-ordered surface is obtained by cleaving the sample in situ using the Scotch tape method at a pressure of 3×10^{-10} mbar followed by a rapid recovery back to the base pressure within 1 min. During the measurements the sample temperature is maintained at 80 K. The Dirac point of the sample is located 200 meV below the Fermi energy.

Reconstruction of p-polarized THz radiation. The electric field component of the p-polarized THz transient is reconstructed from the measured photoelectron energy streaking by considering the classical motion of the photoelectrons in the electric and magnetic fields of the THz transient in vacuum. The transient Lorentz force on a photoemitted electron due to the electric and magnetic field components of the THz transient is

$$\boldsymbol{F}(t) = -e\boldsymbol{E}(t) - e\boldsymbol{v}(t)\times\boldsymbol{B}(t)$$

Here, e is the elementary charge, $\boldsymbol{v}(t)$ the velocity of the electron, and $\boldsymbol{E}(t)$ and $\boldsymbol{B}(t)$ are the electric and magnetic field components of the THz transient. Because the amplitude of the magnetic field component in vacuum is $B_0=E_0/c$ and $v\ll c$, we can neglect the magnetic contribution to the Lorentz force; the velocity of the electron is then

$$\boldsymbol{v}(t) = -\frac{e}{m}\int_t^{\infty}\boldsymbol{E}(t')\mathrm{d}t' + \boldsymbol{v}_0$$

where m is the electron mass, $\boldsymbol{v}_0$ is the initial velocity of the photoemitted electrons and t is the time of photoemission. We note that $\boldsymbol{v}=\boldsymbol{v}_0$ not only when the photoemission UV pulse arrives after the THz transient has already left the surface, but also when the UV pulse precedes the complete THz transient as the time-averaged acceleration of the complete electric field transient on the electron in the vacuum vanishes.

With kinetic energy $\varepsilon_{\mathrm{kin}}(t)=(1/2)m[\boldsymbol{v}(t)]^2$, the temporal change in the kinetic energy caused by the electric field is

$$\frac{\mathrm{d}\varepsilon_{\mathrm{kin}}(t)}{\mathrm{d}t} = -e\boldsymbol{v}(t)\cdot\boldsymbol{E}(t)$$

In the case of normal emission and screening of the electric field component parallel to the surface, this can be rewritten as

$$\frac{\mathrm{d}\varepsilon_{\mathrm{kin}}(t)}{\mathrm{d}t} = -e\sqrt{2m\varepsilon_{\mathrm{kin}}(t)}\,E_{\perp}(t)$$

where $E_{\perp}(t)$ is the magnitude of the standing wave perpendicular to the surface. Thus, we can quantitatively determine the strength of the electric field along the surface normal directly from the observed transient change in the kinetic energy of the electron using

$$E_{\perp}(t) = -\frac{\mathrm{d}\varepsilon_{\mathrm{kin}}(t)}{\mathrm{d}t}[2me^2\varepsilon_{\mathrm{kin}}(t)]^{-1/2}$$

Reconstruction of s-polarized THz radiation. The s-polarized electric field is oriented parallel to the surface, along the direction in which our electron analyser detects the component of the electron momentum parallel to the surface ($\hbar k_y$). Streaking is strongly supressed in this geometry because of the screening of the THz field parallel to the surface. Nevertheless, the photoemitted electrons acquire a small additional parallel momentum

$$\Delta p(t) = -e\int_t^{\infty}E(t')\mathrm{d}t'$$

leading to an overall momentum shift of the photoelectron spectrum. In addition, they experience a small energy shift $\Delta\varepsilon=\Delta p v_{0,\|}$ proportional to their initial parallel velocity $v_{0,\|}$. As a result, the photoelectron spectra of the Dirac cone appear not only shifted in k_y direction but also slightly tilted. This tilt can in principle be used to distinguish the acceleration of electrons by the electric field in the solid (which affects only electrons in the vicinity of the Fermi level) from the acceleration after photoemission (which affects all electrons). In our experiment, the parallel velocity of electrons photoemitted from the Fermi level ($\hbar k_{\mathrm{F}}/m=0.87$ Å fs^{-1}) is much smaller than their velocity in the solid ($v_{\mathrm{F}}=4.1$ Å fs^{-1}). The observed transient energy gains and losses of electrons near k_{F} due to the s-polarized THz field are thus dominated by the acceleration inside the solid and we refrain from applying the corresponding correction to the data.

The small momentum streaking, however, allows us to track the instantaneous THz field in the case of s-polarized THz radiation using

$$E_{\|}(t) = -\frac{\hbar}{e}\frac{\mathrm{d}k_{\mathrm{streak}}}{\mathrm{d}t}$$

where $\mathrm{d}k_{\mathrm{streak}}/\mathrm{d}t$ is the transient momentum shift of the whole photoemission spectrum. At the maximum field amplitude, the observed transient momentum shift was 0.043 ± 0.009 Å^{-1} ps^{-1}, corresponding to a THz field of $E_{\|,\max}=2.8\pm0.5$ kV cm^{-1}. Because the parallel component of the electric field in the surface region of the sample is the same as in the vacuum, momentum streaking provides an accurate and direct way of determining the actual strength of the acceleration field that acts on the electrons in the topological surface state.

This field strength can be verified independently by using the Fresnel coefficients[33] of the Bi_2Te_3 sample and the field extracted for p-polarized pulses. To this end, we first compare p-polarized THz waveforms reflected from Bi_2Te_3 and a gold reference by using electro-optic sampling (see Extended Data Fig. 2a). The ratio of the amplitude spectra of the two transients (Extended Data Fig. 2b) yields the reflection coefficient r_{p} (see Extended Data Fig. 2c, black spheres). r_{p} is rather high over the measured frequency range—similar to metals—and has three minima, which we attribute to phonon resonances[34]. Consequently, we model the underlying dielectric function $\epsilon(\omega)$ with the response of a free-electron plasma and three Lorentzian oscillators:

$$\epsilon(\omega) = \epsilon_{\infty} - \frac{\omega_{\mathrm{p}}^2}{\omega^2+i\omega/\tau} + \frac{A_1}{\omega_1^2-\omega^2-i\omega\gamma_1} + \frac{A_2}{\omega_2^2-\omega^2-i\omega\gamma_2} + \frac{A_3}{\omega_3^2-\omega^2-i\omega\gamma_3} \tag{1}$$

where ϵ_{∞} is the high-frequency limit of the dielectric function, ω_{p} denotes the plasma frequency and τ is the damping of the Drude term. Moreover, the three Lorentz oscillators are characterized by the resonance frequencies ω_1, ω_2 and ω_3, the oscillator strengths A_1, A_2 and A_3, and the damping rates γ_1, γ_2 and γ_3. The values for the phonon frequencies and damping rates are taken from elsewhere[34], and the value for the plasma frequency is taken from the independent transport measurement discussed above. Varying A_1, A_2 and A_3 allows us to fit (Extended Data Fig. 2c, orange curve) the experimentally measured reflection coefficient. Because the p-polarized THz waveform of Fig. 2d recorded by energy streaking is

the superposition of incident and reflected fields[33], r_p allows us to deduce the incident field transient. When we assume the same incident waveform in s-polarization (corrected by a scaling factor of 2 set by the wire-grid polarizers) and compute[33] the transmission coefficient t_s (Extended Data Fig. 2d) from the dielectric function of equation (1), we obtain the internal field in s-polarization. The peak amplitude parallel to the surface is found to be $E_{\parallel,\max}=4.0\pm1.5$ kV cm^{-1}, which indeed lies within the error margins of the value obtained directly by photoelectron streaking (Fig. 4a).

Calculation of the current density. The current density of two-dimensional Dirac electrons is given by the k-space integral

$$\boldsymbol{j}=-\nu e v_{\mathrm{F}}\int\frac{\mathrm{d}^2k}{(2\pi)^2}f(\boldsymbol{k})\hat{\boldsymbol{k}} \tag{2}$$

where $f(\boldsymbol{k})$ is the momentum distribution function, $\hat{\boldsymbol{k}}$ is a unit vector and ν denotes the number of electrons per surface unit cell ($\nu=1$ for a topological surface state). In this expression we have exploited the fact that all electrons travel with the same group velocity, $\boldsymbol{v}(\boldsymbol{k})=\boldsymbol{v}_{\mathrm{F}}=v_{\mathrm{F}}\hat{\boldsymbol{k}}$, owing to the linear dispersion of the Dirac cone. Moreover, neglecting the weak spin-Hall effect, $f(\boldsymbol{k})$ stays symmetric in the k_x direction with an applied field in the y direction. Therefore, accurate values for $\boldsymbol{j}(t)$ can be deduced directly from the experimental data by numerical integration using the measured snapshots of $f(k_y, t)$ and the measured value of the Fermi velocity, $v_{\mathrm{F}}=4.1$ Å fs^{-1}.

An approximate value of the magnitude of $\boldsymbol{j}$ is obtained by considering that the observed shifts of $f(k_y, t)$ in the y direction (Δk), although large on an absolute scale, are small relative to the Fermi momentum k_{F}. Because the widening of f due to scattering is also small compared to k_{F}, evaluation of equation (2) yields $j\approx-ev_{\mathrm{F}}k_{\mathrm{F}}\Delta k/(4\pi)$. With the two-dimensional electron density $n=-k_{\mathrm{F}}^2/(4\pi)$, the approximate current density becomes $j\approx-ev_{\mathrm{F}}n\Delta k/k_{\mathrm{F}}$. With the experimental value of $k_{\mathrm{F}}=0.075$ Å^{-1}, the electron density in the upper part of the Dirac cone of the topological surface state of our sample amounts to $n=4.5\times10^{-4}$ Å$^{-2}=4.5\times10^{12}$ cm^{-2}. For the maximum observed momentum shift of $\Delta k=0.005$ Å^{-1}, a substantial fraction of $2\Delta k/k_{\mathrm{F}}=13\%$ of these electrons carry the current. The current density thus reaches values as high as $j=0.12e$ ps^{-1} Å$^{-1}=2.0$ A cm^{-1}. We anticipate that these femtosecond currents may also become accessible to complementary real-space measurements[27] once subcycle resolution has been reached by these techniques.

Modelling electron scattering using the Boltzmann equation. The Boltzmann equation for electrons in a two-dimensional surface state accelerated by a time-dependent electric field $\boldsymbol{E}(t)$ parallel to the surface has the form

$$\frac{\partial f(\boldsymbol{k},t)}{\partial t}=-\frac{e}{\hbar}\boldsymbol{E}(t)\cdot\boldsymbol{\nabla}_{\boldsymbol{k}}f(\boldsymbol{k},t)+\left[\frac{\partial f(\boldsymbol{k},t)}{\partial t}\right]_{\text{scattering}} \tag{3}$$

The first term on the right describes the redistribution of electrons in k space in the presence of $\boldsymbol{E}(t)$. The second term includes all changes in $f(\boldsymbol{k}, t)$ due to electron scattering within the surface state. Scattering into and out of bulk electronic states as well as electron diffusion from the topological surface state into the bulk are absent under the conditions of our experiment because the acceleration field does not lift the electrons above the bulk bandgap of Bi_2Te_3 (energy gain less than 25 meV).

For the simulation, we compose the scattering term from two phenomenological contributions:

$$\left[\frac{\partial f(\boldsymbol{k},t)}{\partial t}\right]_{\text{scattering}}=-\frac{f(\boldsymbol{k},t)-f_0(\boldsymbol{k})}{\tau_{\mathrm{R}}}-\frac{f(\boldsymbol{k},t)-f(-\boldsymbol{k},t)}{\tau_{k\bar{k}}} \tag{4}$$

Microscopically, the first term describes elastic and inelastic electron scattering within the relaxation-time approximation and equilibrates the accelerated electrons, where

$$f_0(\boldsymbol{k})=\frac{1}{\exp[\beta(\hbar v_{\mathrm{F}}\,|\boldsymbol{k}|-\mu)]+1}$$

is the equilibrium momentum distribution before the arrival of the THz pulse, μ is the chemical potential (Fermi level), $\beta=1/(k_{\mathrm{B}}T)$ is the inverse temperature and τ_{R} is a phenomenological relaxation time. In general, τ_{R} should depend on $\boldsymbol{k}$ because, for example, the phase space for inelastic scattering is a function of the momentum along the Dirac cone. However, the momentum interval around k_{F} in which $f(\boldsymbol{k}, t)$ changes is still sufficiently small for scattering to be approximated by a single parameter τ_{R}. The second, purely elastic scattering term drives the accelerated Dirac fermions into a hot quasi-equilibrium distribution. Here, multiple scattering processes are involved microscopically because direct backscattering from $\boldsymbol{k}$ to $-\boldsymbol{k}$ is strongly supressed in a topological surface state as it would require a spin flip[19]. The time constant $\tau_{k\bar{k}}$ represents an effective phenomenological backscattering time due to multiple individual scattering events.

We first discuss the effect of the relaxation term, neglecting the second term in equation (4). The time-dependent Boltzmann equation (equation (3)) can then be transformed into an ordinary differential equation and solved by direct integration. For small relaxation times τ_{R}, the solution is

$$f(\boldsymbol{k},t)\approx\frac{1}{\exp\{\beta[\hbar v_{\mathrm{F}}\,|\boldsymbol{k}|-\mu-e\boldsymbol{v}_{\mathrm{F}}\cdot\boldsymbol{E}(t)\tau_{\mathrm{R}}]\}+1}$$

In this limit, $f(\boldsymbol{k}, t)$ is just the initial distribution $f_0(\boldsymbol{k})=f(\boldsymbol{k}, 0)$ shifted in energy by an amount proportional to the Fermi velocity v_{F} of the Dirac electrons, the scattering time τ and the momentary strength of the electric field $\boldsymbol{E}(t)$. In k space, this energy shift corresponds to $\Delta k=eE\tau_{\mathrm{R}}/\hbar$. This is the same result as usually derived by considering the stationary case $\partial f(\boldsymbol{k}, t)/\partial t=0$ for a slowly varying electric field and with the assumption that under the presence of the electric field $f(\boldsymbol{k})$ does not deviate much from the unperturbed distribution $f_0(\boldsymbol{k})$ (ref. [35]). Likewise, the shape of $f(\boldsymbol{k}, t)$ does not deviate from $f_0(\boldsymbol{k})$ for large scattering times τ_{R} and is merely shifted by an amount proportional to v_{F} and the time-integrated force acting on the electrons.

In our case, the temporal variation of the electric field $\boldsymbol{E}(t)$ occurs on a timescale similar to the relaxation time τ_{R}. Then the distribution $f(\boldsymbol{k}, t)$ not only is shifted in k space but also attains a shape that deviates from the initial Fermi–Dirac distribution $f_0(\boldsymbol{k})$. However, for not too large electric fields, these deviations remain small (Extended Data Fig. 3a-f). By contrast, we observe experimentally that $f(k_y, t)$ widens substantially for longer delays. The broadening is clearly visible in Fig. 3f, and in our experiment (Extended Data Fig. 4) is most obvious at delay times where Δk goes through zero (Extended Data Fig. 4b). To account for this additional broadening, we add the second scattering term in equation (4). The Boltzmann equation (equation (3)) with this backscattering term alone would be able to reproduce the experimentally observed time-dependent shift in $f(\boldsymbol{k}, t)$. In this model, however, the width of the distribution can only increase but not decrease as a function of time (Extended Data Fig. 3g–l). Experimentally, we observe that for delays exceeding 2.5 ps the width has relaxed to its initial value (Extended Data Fig. 4c). For these reasons both terms in equation (4) are necessary in the most basic model description.

Although the assumptions of our phenomenological model are strong simplifications for excitation far from equilibrium, the approach is able to describe the experimental data well (compare Fig. 3e, f and Fig. 4b). Moreover, the experimental determination of the amplitude and of the exact phase of the accelerating field $\boldsymbol{E}(t)$ by means of momentum streaking puts firm lower bounds on the scattering times. Assuming higher field strengths would require faster scattering for the model to reproduce the measured momentum shift of $f(ky, t)$ and the respective current density $j(t)$. However, the simulated curves then precede the measured ones in time. We achieve best overall agreement between experiment and simulation when a time constant for effective backscattering $\tau_{k\bar{k}}$ is chosen that is of similar magnitude (about 1 ps) to the relaxation time τ_{R}.

Joule heating. The Joule heating of the sample by the generated electron current is extremely small, as can be seen from the following. With a current density of $j=2$ A cm^{-1}, as has been generated in an accelerating field of $E=2.4$ kV cm^{-1}, the energy dissipated by one THz pulse within its duration of $\Delta t=2.5$ ps amounts only to $W=jE\Delta t=0.01$ μJ cm^{-2}. This value is several orders of magnitude lower than the absorbed fluence in a usual optical pump–probe experiment. Even if we neglect heat diffusion and assume that the dissipated heat stays within the topmost quintuple layer of Bi_2Te_3 ($d=1$ nm), the temperature rise associated with this energy is only $\Delta T=W/(dc_{\mathrm{p}})=0.1$ K, where we assume $c_{\mathrm{p}}=1.001$ J cm^{-3} K^{-1} as the volumetric heat capacity of Bi_2Te_3 at 80 K, calculated from the specific heat of 103.58 J mol^{-1} K^{-1} (ref. [36]), a molar mass of $M=800.761$ g mol^{-1} and the density of $\rho=7.74$ g cm^{-3}.

Calculation of the current excursion within the topologically protected surface band. From the two-dimensional current density j and the scattering times τ_{R} and $\tau_{k\bar{k}}$ of the Dirac fermions, we can retrieve the microscopic evolution of the current density within the surface layer of Bi_2Te_3. In a first step, we calculate the total scattering time as

$$\tau_{\mathrm{t}}=\left(\frac{1}{\tau_{\mathrm{R}}}+\frac{2}{\tau_{k\bar{k}}}\right)^{-1}$$

Dirac fermions propagating at a velocity of $v_{\mathrm{F}}=4.1$ Å fs^{-1} will then have a mean free path of $l=v_{\mathrm{F}}\tau_{\mathrm{t}}=137$ nm. According to the scattering-time approximation (equation (4)), a current created at location $y=0$, would then decay exponentially with a decay constant of 137 nm. To visualize this situation, we consider the accelerating action of the THz electric field only at $y=0$ and let the current propagate freely elsewhere. With the current density $j(t)=j(y=0, t)$ (Fig. 4b, red curve) extracted from the Boltzmann equation (see equation (3)), we can calculate the full spatial and temporal dynamics of the current density $j(y, t)$ as

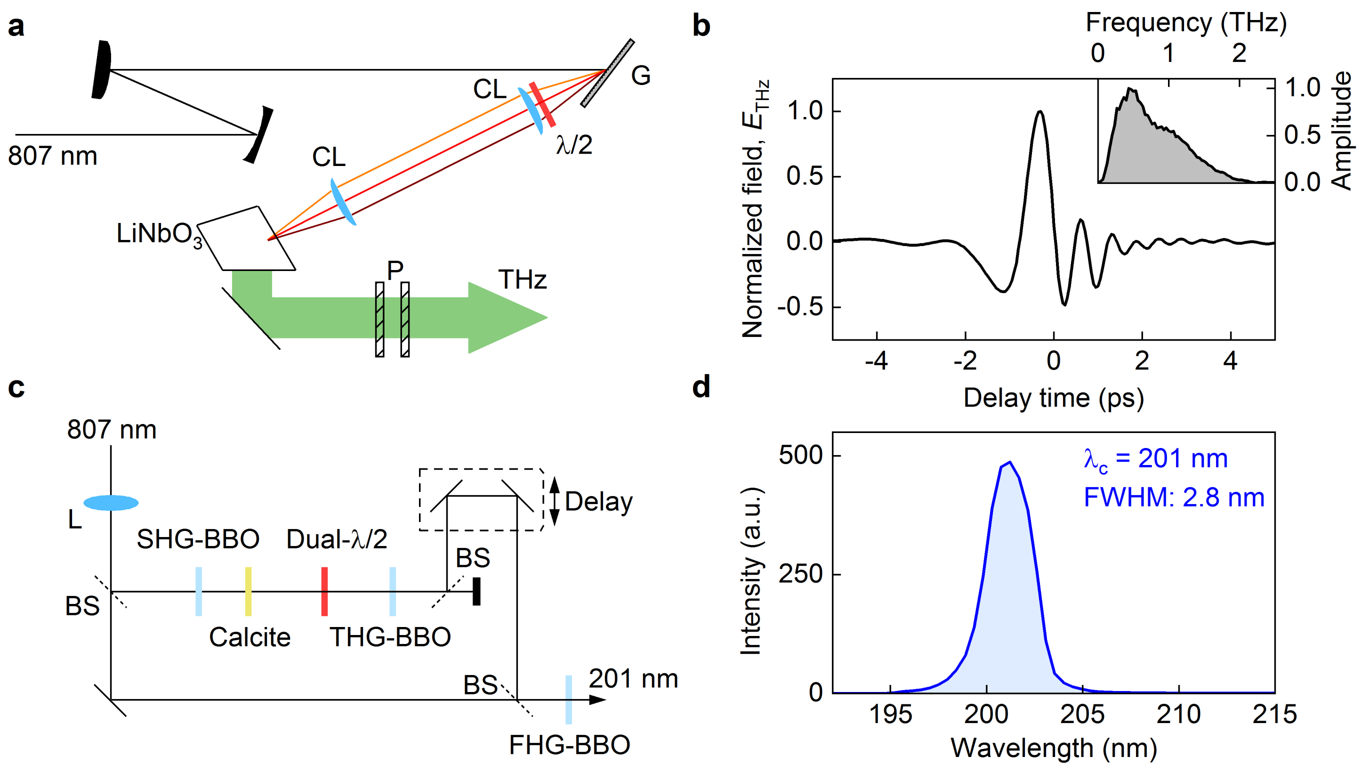
$$j(y,t)=j_0\left(t-\frac{y}{v_{\mathrm{F}}}\right)\mathrm{e}^{-y/l}$$

Extended Data Fig. 5a displays $j(y, t)$ (colour scale) in the surface layer of Bi_2Te_3. A sizable fraction of the initial current density survives even after a propagation length as large as 0.5 μm.

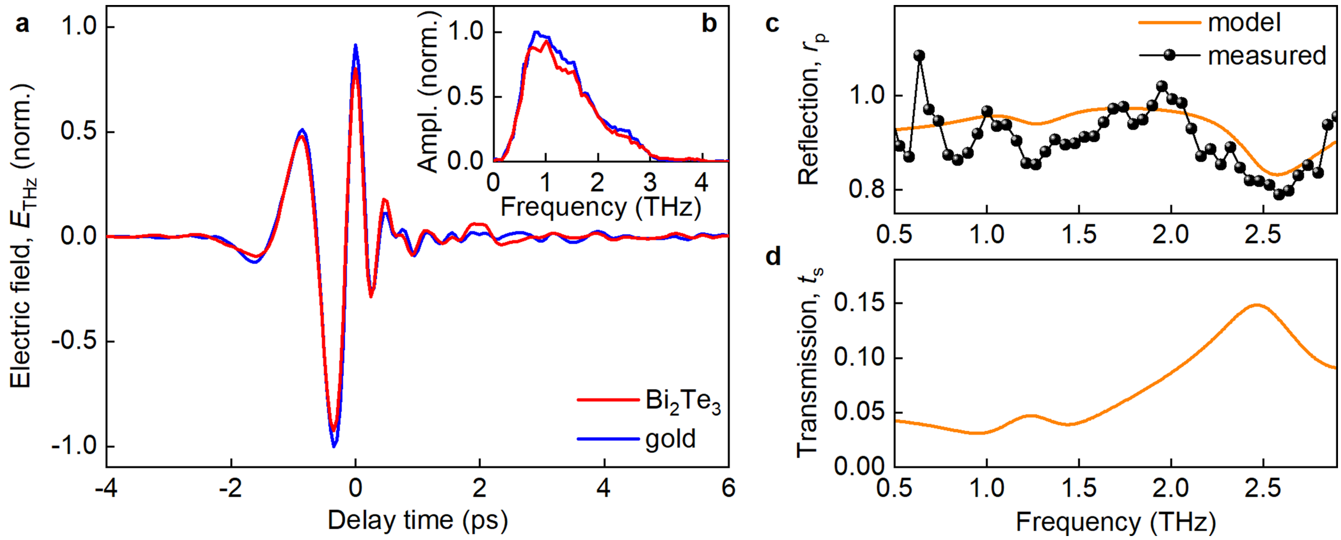
Data availability

The data that support the findings of this study are available from the corresponding authors on request.

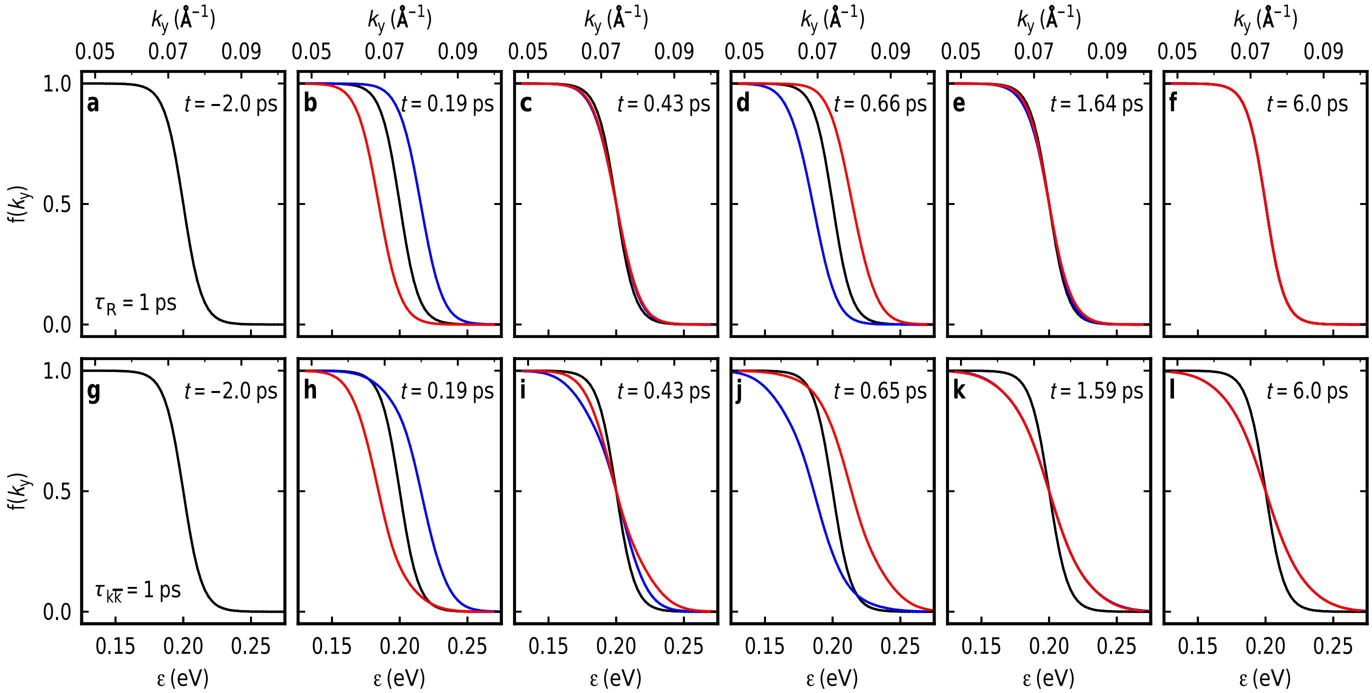
31. Hebling, J., Yeh, K.-L., Hoffmann, M. C., Bartal, B. & Nelson, K. A. Generation of high-power terahertz pulses by tilted-pulse-front excitation and their application possibilities. *J. Opt. Soc. Am. B* **25**, 6–19 (2008).
32. Kokh, K. A. et al. Melt growth of bulk Bi_2Te_3 crystals with a natural p-n junction. *CrystEngComm* **16**, 581–584 (2014).
33. Hecht, E. *Optics* 4th edn, Ch. 4 (Pearson Addison-Wesley, New York, 2012).
34. Richter, W., Köhler, H. & Becker, C.R. Raman and far-infrared investigation of phonons in the rhombohedral V_2-VI_3 compounds Bi_2Te_3, Bi_2Se_3, Sb_2Te_3 and $Bi_2(Te_{1-x}Se_x)_3$ $(0 < x < 1)$, $(Bi_{1-y}Sb_y)_2Te_3$ $(0 < y < 1)$. *Phys. Status Solidi B* **84**, 619–628 (1977).
35. Ziman, J. M. *Principles of the Theory of Solids* 2nd edn, Ch. 7 (Cambridge Univ. Press, Cambridge, 1979).
36. Gorbachuk, N. P. & Sidorko, V. R. Heat capacity and enthalpy of Bi_2Si_3 and Bi_2Te_3 in the temperature range 58–1012 K. *Powder Metall. Met. Ceramics* **43**, 284–290 (2004).



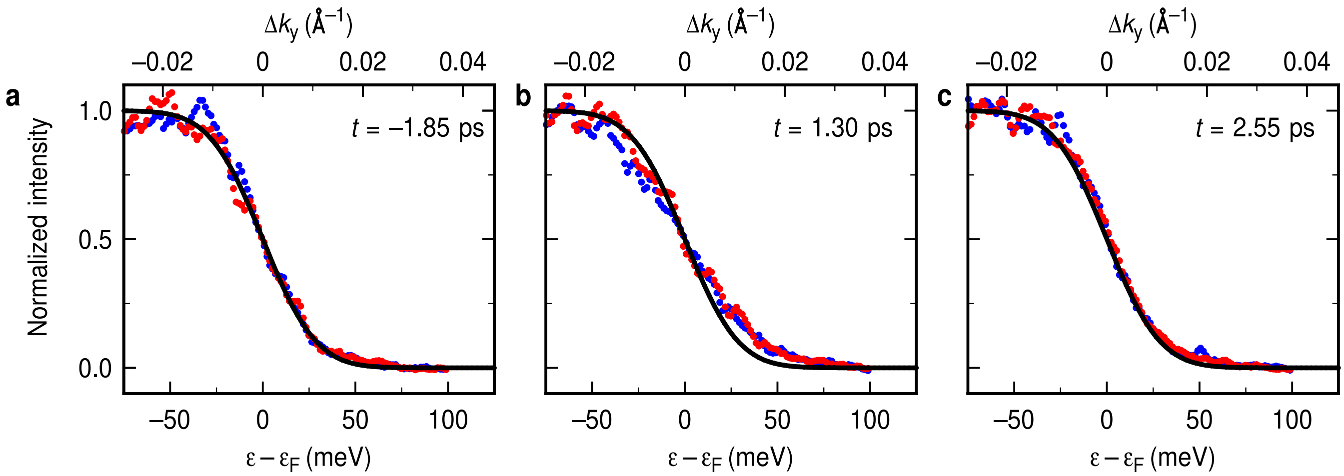
Extended Data Fig. 1 | Optical set-up. a, THz generation via tilted pulse fronts in lithium niobate ($LiNbO_3$). After the laser beam from the titanium:sapphire amplifier system has been reduced in diameter by a reflective telescope, a grating (G) induces a pulse front tilt. Cylindrical lenses (CL) image and focus the beam into a cryogenically cooled $LiNbO_3$ crystal, where optical rectification generates intense THz radiation. A pair of wire-grid polarizers (P) controls the polarization state. **b**, Electro-optically detected THz waveform after transmission through the fused silica window of the vacuum chamber. Inset, amplitude spectrum of the transient shown in **a**. **c**, Set-up for generating UV pulses for photoemission. A lens (L; focal length, 1 m) gently focuses the fundamental titanium:sapphire pulses. A beam splitter (BS) separates 90% of the intensity for second-harmonic generation in a BBO crystal (SHG-BBO). Subsequent dispersion and polarization control is employed using a birefringent calcite plate and a dual-half-wave plate (Dual-λ/2). The third harmonic is generated in another BBO crystal (THG-BBO), separated from the fundamental pulses with a beam splitter (BS), and spatially and temporally (Delay) overlapped with the remaining fundamental pulses to generate the sum-frequency at the fourth harmonic (FHG-BBO). **d**, The resulting spectrum of the fourth harmonic is centred at $\lambda_c = 201$ nm with a full-width at half-maximum (FWHM) of 2.8 nm, corresponding to a Fourier limit of 22 fs.



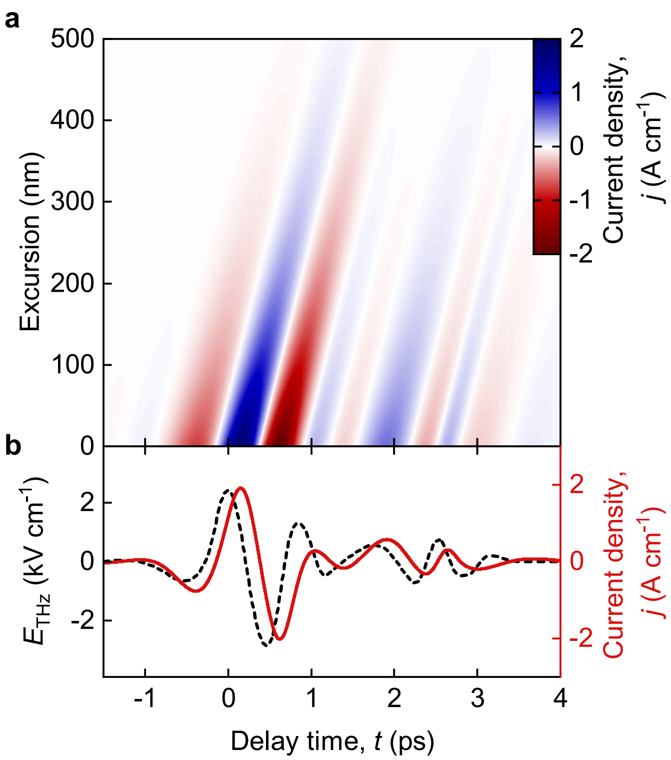
Extended Data Fig. 2 | Determination of the reflectivity of Bi_2Te_3. **a**, Electro-optically detected THz transients after reflection off a Bi_2Te_3 (red) and a gold reference (blue) surface kept at a temperature of 77 K. **b**, Amplitude spectra of the field traces in **a** normalized to the amplitude of the gold spectrum. **c**, Reflection coefficient r_p (black spheres) for parallel incidence obtained by dividing the amplitude spectra from Bi_2Te_3 and gold. The orange curve describes the reflection coefficient r_p calculated using the corresponding Fresnel formula with a Drude–Lorentz model for the dielectric function of Bi_2Te_3 (see equation (1)). **d**, Transmission coefficient t_s calculated using the modelled dielectric function.



Extended Data Fig. 3 | Comparison of scattering mechanisms. **a–l**, Calculated distribution functions at different delay times *t* for the experimental THz waveform with an amplitude of 2.4 kV cm^{-1}. Experimental broadening is not accounted for. The Boltzmann equation used to compute these results includes only the relaxation term (**a–f**) or only the effective backscattering term (**g–l**).



Extended Data Fig. 4 | Broadening of the electron distribution. a–c, Experimental distribution functions (red and blue circles) for different delay times t compared to the equilibrium Fermi–Dirac distribution (solid black line).



Extended Data Fig. 5 | Local current density in Bi_2Te_3. a, Calculated current density *j* (colour scale) as a function of the delay time *t* and the electron excursion in the surface layer of Bi_2Te_3. The excursion was evaluated using the extracted scattering times of the charge carriers within the topologically protected surface state. The intense THz fields coherently drive Dirac fermions over several hundred nanometres before they undergo scattering. **b**, The red solid curve shows the simulated current dynamics calculated for scattering times of $\tau_R = \tau_{k\bar{k}} = 1$ ps; the THz electric field is depicted as a dashed black curve.

https://doi.org/10.1038/s41586-018-0598-9

Battery-operated integrated frequency comb generator

Brian Stern[1,2], Xingchen Ji[1,2], Yoshitomo Okawachi[3], Alexander L. Gaeta[3] & Michal Lipson[2]*

Optical frequency combs are broadband sources that offer mutually coherent, equidistant spectral lines with unprecedented precision in frequency and timing for an array of applications[1]. Frequency combs generated in microresonators through the Kerr nonlinearity require a single-frequency pump laser and have the potential to provide highly compact, scalable and power-efficient devices[2,3]. Here we demonstrate a device—a laser-integrated Kerr frequency comb generator—that fulfils this potential through use of extremely low-loss silicon nitride waveguides that form both the microresonator and an integrated laser cavity. Our device generates low-noise soliton-mode-locked combs with a repetition rate of 194 gigahertz at wavelengths near 1,550 nanometres using only 98 milliwatts of electrical pump power. The dual-cavity configuration that we use combines the laser and microresonator, demonstrating the flexibility afforded by close integration of these components, and together with the ultra low power consumption should enable production of highly portable and robust frequency and timing references, sensors and signal sources. This chip-based integration of microresonators and lasers should also provide tools with which to investigate the dynamics of comb and soliton generation.

Frequency combs based on chip-scale microresonators offer the potential for high-precision photonic devices for time and frequency applications using a highly compact and robust platform. By pumping the microresonator with a single-frequency pump laser, additional discrete, equidistant frequencies are generated through parametric four-wave mixing (FWM), resulting in a Kerr frequency comb[2,4–7]. Under suitable conditions temporal cavity solitons can be excited, which results in stable, low-noise combs with ultraprecise spacing[8–13]. Many applications require such tight frequency and timing stability, including spectroscopy[14–16], low-noise microwave generation[17], photonic frequency synthesis[18], optical clocks[19], distance ranging[20,21] and telecommunications[22].

Although one of the most compelling advantages for microresonator combs is the potential for the pump source and the microresonator to be fully integrated, previous demonstrations using integrated resonators have relied on external pump lasers that are typically large, expensive and power-hungry, preventing applications where size, portability and low power consumption are critical. Power-efficient integrated lasers have been developed using silicon laser cavities with bonded or attached III–V materials to provide optical gain[23–26], but losses in these silicon waveguides make comb generation impractical at low power. On the other hand, silicon nitride (Si_3N_4) microresonators were recently demonstrated with record low parametric oscillation thresholds[27] due to the high quality factors ($Q > 3 \times 10^7$), high nonlinear refractive index ($n_2 \approx 2.4 \times 10^{-19}$ m^2 W^{-1}) and small mode volume (ring radius approximately 100 μm). Additionally, owing to the high index of refraction of Si_3N_4 ($n \approx 2.0$) and its low loss, compact, tunable Si_3N_4 laser cavities with narrow linewidth have been demonstrated[28,29]. Si_3N_4 is a common complementary metal oxide semiconductor (CMOS)-compatible deposited material that can be fabricated at wafer scale, and the combination of efficient comb generation and available integration of active devices make it an ideal platform for complete integration of optical frequency combs.

Here we demonstrate a Kerr comb source on an integrated hybrid III–V/Si_3N_4 platform, using a compact, low-power, electrically pumped source. In our approach (Fig. 1), a gain section based on a III–V reflective semiconductor optical amplifier (RSOA) is coupled to a Si_3N_4 laser cavity, which consists of two Vernier microring filters for wavelength tunability and a high-Q nonlinear microresonator (Fig. 1b). The nonlinear microresonator serves two purposes. First, it generates a narrowband back-reflection due to coupling between counter-propagating circulating beams resulting from Rayleigh scattering[30], effectively serving as an output mirror of the pump laser cavity, as we previously demonstrated[28]. Second, the microresonator generates a frequency comb through parametric FWM. In this way, the comb generation and pump laser are inherently aligned, a configuration that was previously explored using resonators in fibre laser cavities with fibre amplifiers[31,32]. Integrating the comb source with the laser allows the flexibility to use such a configuration, avoiding the typical chain of discrete components found in all previous Kerr comb demonstrations. Figure 1c shows the assembled millimetre-sized comb source, which has only electrical inputs and an optical output (see Methods for fabrication details).

We designed the Si_3N_4 laser cavity to ensure tunable, single-mode lasing and provide sufficient pump output power for comb generation in the nonlinear microresonator. The lasing wavelength is controlled by the alignment of the two microring Vernier filters[25], which are in turn aligned with one of the modes of the larger microresonator shown in Fig. 1b. The filters' radii are 20 μm and 22 μm, corresponding to free spectral ranges (FSRs) of 1.18 THz and 1.07 THz, respectively, which result in transmission at only a single frequency when the filters are aligned. Their resonance positions can be widely tuned using integrated resistive microheaters, as shown in Fig. 2a. The filters' transmission bandwidth is designed to have a full-width at half-maximum of 15 GHz by ensuring strong coupling to the two adjacent waveguides with a 5 μm coupling length. The optical gain in the laser cavity comes from electrical pumping of the III–V waveguide on the RSOA, which is coupled to the Si_3N_4 cavity at one end and strongly reflects at the opposite end (see Methods). The output coupler of the laser cavity is a 120-μm-radius microresonator with a measured reflection of 40% on resonance, as shown in Fig. 2b. This level of reflection allows for high laser output power due to the high round-trip gain of the RSOA. The measured transmission spectrum of the microresonator (Fig. 2b) corresponds to an intrinsic Q of $(8.0 \pm 0.8) \times 10^6$. Based on this Q and the anomalous group-velocity dispersion for the 730 nm× 1,800 nm waveguide, simulations indicate that a soliton-state frequency comb can be generated with 700 μW of pump power in the bus waveguide just before the microresonator (Extended Data Fig. 1).

We find lasing with up to 9.5 mW output optical power using the integrated Si_3N_4 laser. In order to characterize the laser, we first operate the microresonator slightly detuned from resonance to ensure that only

[1]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA. [2]Department of Electrical Engineering, Columbia University, New York, NY, USA. [3]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA. *e-mail: ml3745@columbia.edu

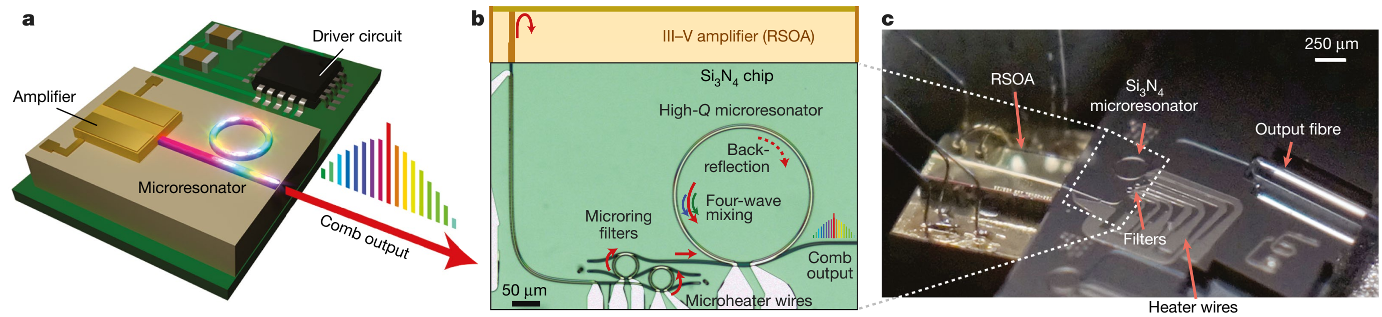


Fig. 1 | Integrated frequency comb source. a, The concept of an integrated Kerr comb source with an on-chip amplifier and microresonator. **b**, Microscope image and diagram of the integrated comb source, including the laser cavity and the high-*Q* nonlinear microresonator for comb generation. The reflective III–V semiconductor optical amplifier (RSOA) waveguide provides electrically pumped optical gain and includes a reflective facet on one end (top), while the opposite side is coupled to the Si_3N_4 portion of the laser cavity. The microring filters and the larger microresonator are tunable using integrated microheaters. The latter generates a partially reflected beam to form a second effective mirror of the laser cavity. This microresonator also has a high *Q* to enable FWM and comb generation. **c**, Photograph of the integrated comb source. The RSOA is edge-coupled to the Si_3N_4 chip and supplied with electric current via wires, while the comb output is measured using an optical fibre.

lasing occurs and a frequency comb is not generated. We observe lasing with a side-mode suppression ratio (SMSR) of more than 60 dB (Fig. 2c). As shown in Fig. 2d, the lasing threshold is 49 mA, with a slope efficiency of 52 mW A^{-1}. The maximum on-chip output power of 9.5 mW is obtained at 277 mW (220 mA) electrical pump power consumption, P_{elec}. This corresponds to a 3.4% wall-plug efficiency (that is, output optical power divided by electrical power). Additionally, we measure a narrow laser linewidth of 40 kHz using the delayed self-heterodyne method (see Methods). The relatively high output power and narrow linewidth are competitive with those of many bulk pump lasers, yet the present laser is much more compact.

Using our cavity design, we generate a Kerr comb spanning more than 8 THz and achieve a mode-locked, single-soliton state with P_{elec} less than 100 mW, enabling battery-operation applications. We observe new optical frequencies beginning to appear adjacent to the 1,579 nm pump owing to FWM in the microresonator once the laser power measured after the ring (P_{opt}) reaches a threshold of 1.1 mW at $P_{elec} = 78$ mW. We then increase P_{elec} above threshold to 130 mW and monitor comb formation as the microresonator is tuned using its integrated microheater (see Methods for the set-up and tuning procedure). When the microresonator is first detuned slightly, we measure $P_{opt} = 2.5$ mW for the single lasing mode (Fig. 3a). As the microresonator is tuned into resonance, greater circulating power leads to comb

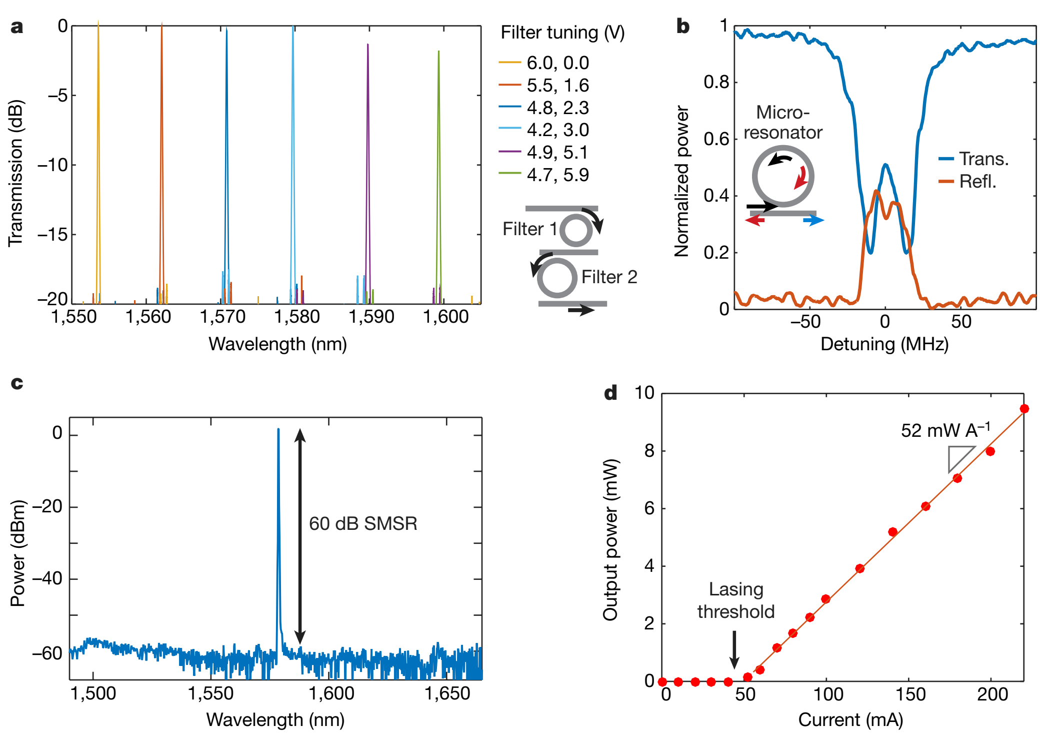


Fig. 2 | Characterization of the integrated III–V/Si_3N_4 laser. a, Measured transmission spectra (normalized) for the Vernier filter microrings (filter 1 and filter 2, diagram at right). By adjusting the voltage applied to the microheaters, the filters' relative detuning is adjusted and a single transmission wavelength is selected. Key at right shows voltage applied in the format 'filter 1, filter 2'. **b**, Measured optical transmission and reflection spectra (normalized) of the high-*Q* microresonator. The 32-MHz resonance bandwidth reveals a *Q* of 8×10^6. The narrowband reflection is generated by coupling via Rayleigh scattering between counter-propagating beams in the ring (arrows show beam directions, colour code as spectra), which is apparent due to the resonance splitting observed from these degenerate beams. **c**, Laser output spectrum at 85 mA showing single-mode lasing with a side-mode suppression ratio (SMSR) of more than 60 dB. **d**, Output optical power of laser versus pump current at 1,580 nm with a slope efficiency of 52 mW A^{-1}.

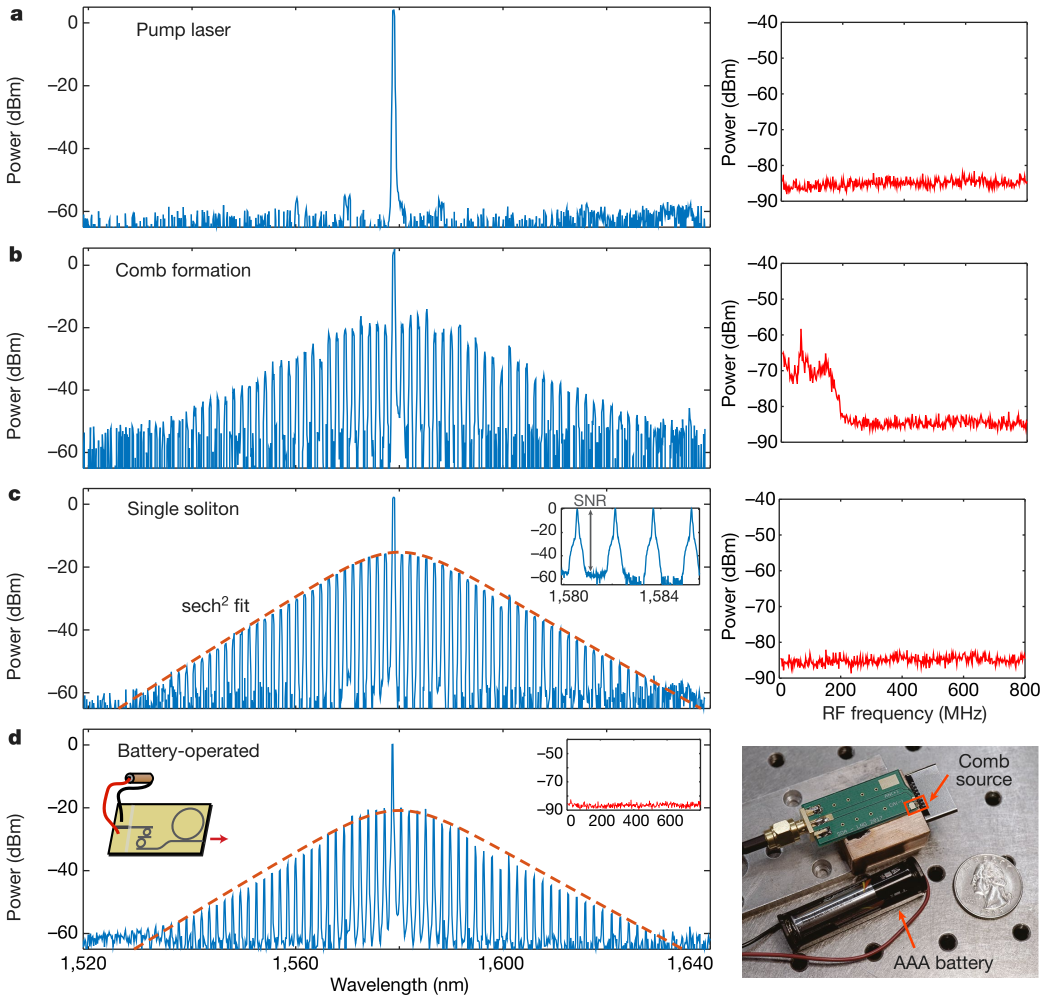


Fig. 3 | Generation of mode-locked soliton frequency combs. **a–c**, Left, spectra of output from the comb source as measured by an optical spectrum analyser at varying stages of comb generation; right, corresponding RF spectra (resolution bandwidth 100 kHz). A current supply provides electrical pump power of 130 mW. **a**, Spectrum of the laser output before tuning fully into resonance. The RF noise is low since there is only single-frequency lasing. **b**, Spectrum of the frequency comb. Because the comb is not yet mode-locked, beating between different comb lines produces high RF noise below 200 MHz. **c**, Single-soliton frequency comb spectrum with the characteristic sech profile (see Extended Data Fig. 2). Inset, the signal-to-noise ratio (SNR, grey vertical arrow) is approximately 50 dB; variables plotted on the axes are the same as in the main panel. The comb linewidth is separately measured as 40 kHz (see Methods). The RF spectrum confirms the transition to a low-noise state. **d**, Left panel, frequency comb spectrum matching soliton sech profile generated with an AAA battery supplying pump power of 98 mW. Left inset, diagram of battery operated device, showing filters 1 and 2 (see Fig. 2a) and the microresonator (large circle). Right inset, RF spectrum showing low-noise state; *y* axis, power in dBm, *x* axis, RF frequency in MHz. Right panel, photograph of integrated comb source (shown boxed in red) with a printed circuit board and the battery (arrowed) next to a US quarter for scale.

formation, accompanied by high radio frequency (RF) noise (Fig. 3b). Tuning the resonance further results in stable combs with smooth spectral envelopes characteristic of temporal cavity solitons[8]. We measure a single-soliton state with a 8.6 THz (72 nm) 30-dB bandwidth accompanied by a drop in RF noise (Fig. 3c). Once generated, the soliton exhibits stable behaviour without feedback electronics or temperature control, with no visible changes in the optical spectrum or output power until the microresonator is intentionally detuned. The power of the comb lines totals 0.24 mW, indicating that a higher effective pump power may be resulting from our placement of the microresonator in the laser cavity. Such efficient operation allows us to also show battery-operation of the comb source by supplying the pump current using a standard AAA battery. At $P_{elec} = 98$ mW from the battery, we measure $P_{opt} = 1.3$ mW and a comb matching the single-soliton profile (Fig. 3d). These results represent unprecedented low-power consumption for generating Kerr frequency combs and solitons with an integrated microresonator.

In order to show the versatility of this platform, we demonstrate a more traditional but still laser-integrated configuration in which the comb is generated in a microresonator that is distinct from the pump laser. In this second design, shown in Fig. 4a, the Vernier filters and RSOA function in the same way as in the first design, but a Sagnac loop mirror is now included to serve as the output coupler with approximately 20% reflection. Because this mirror has a broadband reflection, tunable lasing can take place independent of the resonance position of the comb microresonator. With the microresonator fully off-resonance, we measure single-mode lasing at 1,582 nm with $P_{opt} = 4.9$ mW and over 60 dB SMSR (Fig. 4b) at $P_{elec} = 162$ mW. By tuning the microresonator into resonance with the laser wavelength, we can generate a frequency comb (Fig. 4c). Through further tuning of the resonance (see Methods), we observe a multiple-soliton-state frequency comb spanning a 13.4 THz (105 nm) 30-dB bandwidth with the characteristic drop in RF noise (Fig. 4d). We model a two-soliton-state comb and obtain a profile closely matching that of the experimental comb (Fig. 4d). Single-soliton combs should also be achievable with this configuration, but in this device we only observed two or more solitons. Multiple-soliton combs in microresonators have been used previously to demonstrate dual-comb spectroscopy, for example[16]. The measured comb power is 80 μW, corresponding to a conversion efficiency of 1.6%. The comb power scales with the number of solitons, as does the number

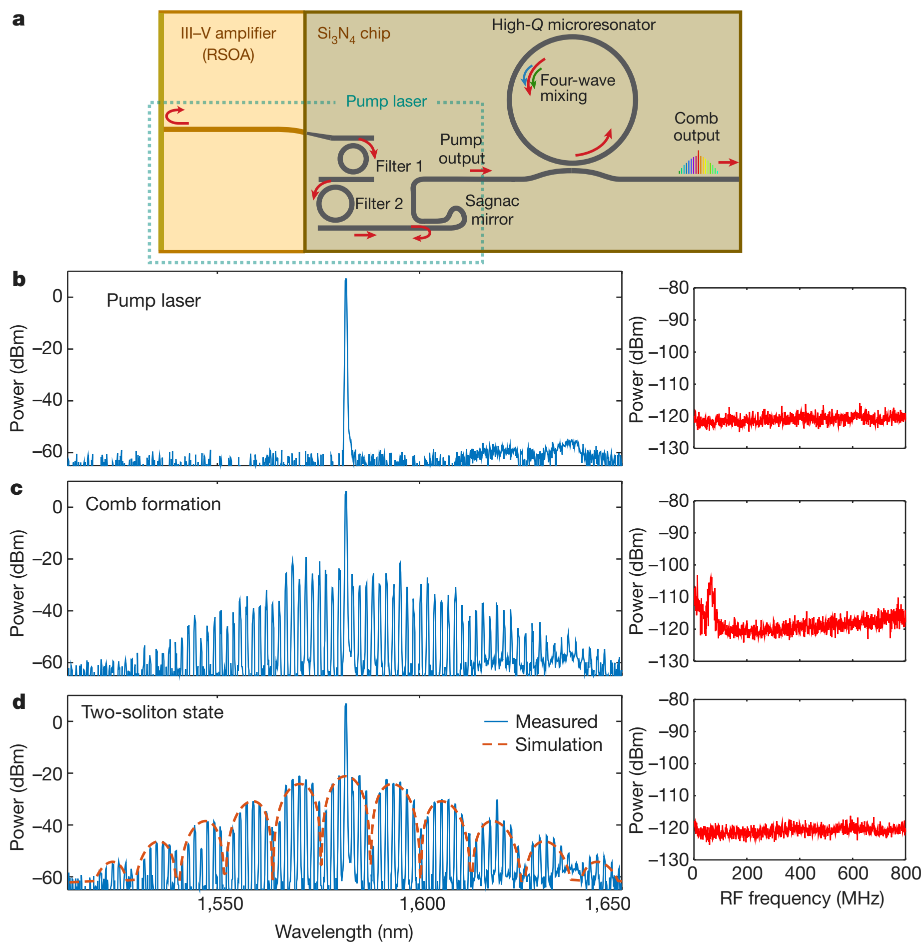


Fig. 4 | Modular configuration of the integrated comb source. **a**, Schematic of the modular comb source configuration. Here the integrated laser (turquoise dashed box) is distinct from the nonlinear microresonator, with a Sagnac loop mirror serving as the laser output coupler. The arrows show the path of light travelling through the laser cavity and reflecting back at the reflective end (left) of the RSOA and at the Sagnac mirror, with the laser output partially transmitting through the latter. **b**–**d**, Optical output spectra at varying stages of comb generation (left panel) with corresponding RF spectra (right panel; resolution bandwidth 100 kHz). **b**, Spectrum of laser output. The RF noise is low because there is only single-frequency lasing. **c**, Spectrum of frequency comb before mode-locking with associated high RF noise. **d**, Spectrum of two-soliton frequency comb. The RF spectrum confirms the low-noise state.

of pulses per round-trip. In Methods, we discuss the relative advantages of the two designs.

This demonstration of a laser-integrated Kerr comb source presents opportunities in many fields that rely on the precision and stability of frequency combs and solitons, including sensing, metrology, communications and waveform generation. The low power consumption of our platform enables these applications in a battery-powered and mobile system without the need for external lasers, moveable optics, or laboratory set-ups. Our platform is CMOS-compatible for wafer-scale fabrication of robust integrated photonic chips, potentially enabling wide deployment of precision devices, such as portable spectrometers for molecular sensing[14,15] or vehicle-mounted systems for distance ranging[20,21]. In future implementations, the RSOA could be placed directly on the silicon substrate, through passively aligned mounting[25] or material bonding[23], taking advantage of the infrastructure for assembly and packaging of III–V and silicon chips that is already scaled to mass production for silicon photonic transceivers. Additional photonic components such as filters for wavelength-division multiplexing[22] or waveguide couplers for mixing multiple combs[14,15,18] could also be placed on-chip to combine frequency combs with more complex integrated photonic circuits.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0598-9.

Received: 30 March 2018; Accepted: 8 August 2018;
Published online 8 October 2018.

1. Newbury, N. R. Searching for applications with a fine-tooth comb. *Nat. Photon.* **5**, 186–188 (2011).
2. Del'Haye, P. et al. Optical frequency comb generation from a monolithic microresonator. *Nature* **450**, 1214–1217 (2007).
3. Pasquazi, A. et al. Micro-combs: a novel generation of optical sources. *Phys. Rep.* **729**, 1–81 (2018).
4. Jung, H., Xiong, C., Fong, K. Y., Zhang, X. & Tang, H. X. Optical frequency comb generation from aluminum nitride microring resonator. *Opt. Lett.* **38**, 2810–2813 (2013).
5. Savchenkov, A. A. et al. Tunable optical frequency comb with a crystalline whispering gallery mode resonator. *Phys. Rev. Lett.* **101**, 093902 (2008).
6. Levy, J. S. et al. CMOS-compatible multiple-wavelength oscillator for on-chip optical interconnects. *Nat. Photon.* **4**, 37–40 (2010).
7. Razzari, L. et al. CMOS-compatible integrated optical hyper-parametric oscillator. *Nat. Photon.* **4**, 41–45 (2010).
8. Herr, T. et al. Temporal solitons in optical microresonators. *Nat. Photon.* **8**, 145–152 (2014).
9. Saha, K. et al. Modelocking and femtosecond pulse generation in chip-based frequency combs. *Opt. Express* **21**, 1335–1343 (2013).
10. Yi, X., Yang, Q.-F., Yang, K. Y., Suh, M.-G. & Vahala, K. Soliton frequency comb at microwave rates in a high-Q silica microresonator. *Optica* **2**, 1078–1085 (2015).
11. Yu, M., Okawachi, Y., Griffith, A. G., Lipson, M. & Gaeta, A. L. Mode-locked mid-infrared frequency combs in a silicon microresonator. *Optica* **3**, 854–860 (2016).
12. Xue, X. et al. Mode-locked dark pulse Kerr combs in normal-dispersion microresonators. *Nat. Photon.* **9**, 594–600 (2015).
13. Volet, N. et al. Micro-resonator soliton generated directly with a diode laser. *Laser Photonics Rev.* **12**, 1700307 (2018).

14. Suh, M.-G., Yang, Q.-F., Yang, K. Y., Yi, X. & Vahala, K. J. Microresonator soliton dual-comb spectroscopy. *Science* **354**, 600–603 (2016).
15. Dutt, A. et al. On-chip dual-comb source for spectroscopy. *Sci. Adv.* **4**, e1701858 (2018).
16. Yu, M. et al. Silicon-chip-based mid-infrared dual-comb spectroscopy. *Nat. Commun.* **9**, 1869 (2018).
17. Liang, W. et al. High spectral purity Kerr frequency comb radio frequency photonic oscillator. *Nat. Commun.* **6**, 7957 (2015).
18. Spencer, D. T. et al. An optical-frequency synthesizer using integrated photonics. *Nature* **557**, 81–85 (2018).
19. Papp, S. B. et al. Microresonator frequency comb optical clock. *Optica* **1**, 10–14 (2014).
20. Suh, M.-G. & Vahala, K. J. Soliton microcomb range measurement. *Science* **359**, 884–887 (2018).
21. Trocha, P. et al. Ultrafast optical ranging using microresonator soliton frequency combs. *Science* **359**, 887–891 (2018).
22. Marin-Palomo, P. et al. Microresonator-based solitons for massively parallel coherent optical communications. *Nature* **546**, 274–279 (2017).
23. Fang, A. W. et al. Electrically pumped hybrid AlGaInAs-silicon evanescent laser. *Opt. Express* **14**, 9203–9210 (2006).
24. Van Campenhout, J. et al. Electrically pumped InP-based microdisk lasers integrated with a nanophotonic silicon-on-insulator waveguide circuit. *Opt. Express* **15**, 6744–6749 (2007).
25. Kobayashi, N. et al. Silicon photonic hybrid ring-filter external cavity wavelength tunable lasers. *J. Lightwave Technol.* **33**, 1241–1246 (2015).
26. Lee, J.-H. et al. Demonstration of 12.2% wall plug efficiency in uncooled single mode external-cavity tunable Si/III-V hybrid laser. *Opt. Express* **23**, 12079–12088 (2015).
27. Ji, X. et al. Ultra-low-loss on-chip resonators with sub-milliwatt parametric oscillation threshold. *Optica* **4**, 619–624 (2017).
28. Stern, B., Ji, X., Dutt, A. & Lipson, M. Compact narrow-linewidth integrated laser based on a low-loss silicon nitride ring resonator. *Opt. Lett.* **42**, 4541–4544 (2017).
29. Oldenbeuving, R. M. et al. 25 kHz narrow spectral bandwidth of a wavelength tunable diode laser with a short waveguide-based external cavity. *Laser Phys. Lett.* **10**, 015804 (2013).
30. Liang, W. et al. Whispering-gallery-mode-resonator-based ultranarrow linewidth external-cavity semiconductor laser. *Opt. Lett.* **35**, 2822–2824 (2010).
31. Pasquazi, A. et al. Self-locked optical parametric oscillation in a CMOS compatible microring resonator: a route to robust optical frequency comb generation on a chip. *Opt. Express* **21**, 13333–13341 (2013).
32. Johnson, A. R. et al. Microresonator-based comb generation without an external laser source. *Opt. Express* **22**, 1394–1401 (2014).

Acknowledgements We are grateful to S. Miller, C. Joshi, T. Lin, U. Dave and J. Jang for discussions and to M. Yu for help with soliton simulations. We also thank M. C. Shin and O. Jimenez for packaging advice. This work was supported by AFRL programme award number FA8650-17-P-1085; the ARPA-E ENLITENED programme (DE-AR0000843); the Defense Advanced Research Projects Agency (DARPA) under the Microsystems Technology Office Direct On-Chip Digital Optical Synthesizer (DODOS) program (N66001-16-1-4052) and the Modular Optical Aperture Building Blocks (MOABB) programme (HR0011-16-C-0107); the STTR programme (N00014-16-P-30); and the Air Force Office of Scientific Research (AFOSR) (FA9550-15-1-0303). X.J. acknowledges the China Scholarship Council for financial support. This work was performed in part at the Cornell NanoScale Facility, an NNCI member supported by NSF grant ECCS-1542081.

Reviewer information *Nature* thanks W. Freude and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.S. conceived the work, designed and assembled the devices, performed the measurements, and prepared the manuscript. X.J. fabricated the devices. B.S. and X.J. characterized the microring transmission. Y.O. simulated the soliton combs. M.L. and A.L.G. supervised the project. All authors discussed the results and edited the manuscript.

Competing interests All authors are listed as inventors in a patent application related to this work, filed by Columbia University.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0598-9.

Correspondence and requests for materials should be addressed to M.L.

METHODS

Fabrication. The Si_3N_4 devices are fabricated[27] by first growing 4 μm of SiO_2 on a crystalline silicon wafer using thermal oxidation to form the bottom cladding of the waveguides. Then 730 nm of Si_3N_4 is deposited using low pressure chemical vapour deposition (LPCVD). The wafer is annealed in two stages to remove hydrogen impurities. The waveguides are then patterned using electron beam lithography and etched using CHF_3 plasma etching. The waveguides are clad with 2 μm SiO_2. The microheaters are placed over the waveguides using 100 nm of sputtered platinum (with a titanium adhesion layer) and lift-off patterning.

RSOA/Si_3N_4 coupling and electrical connection. The III–V RSOA gain chip used here is commercially available from Thorlabs (SAF 1126) and provides broad gain near 1,550 nm. One side has 93% reflection and the other side is anti-reflection coated. This second side is coupled to the Si_3N_4 chip with the waveguides angled relative to the facets to further prevent reflections[28]. The Si_3N_4 chip is polished up to the end of a tapered 280-nm-wide waveguide which is simulated to have less than 1 dB coupling loss to the mode of the RSOA waveguide. The two chips are attached and aligned using three-axis stages with micrometers. We measure an experimental 2 dB coupling loss. The RSOA is wirebonded to an electrical printed circuit board (PCB) for supplying the pump current from either a Keithley 2400 SourceMeter or an AAA battery with a tunable potentiometer. The microheaters of the Si_3N_4 chip are connected to pads and interfaced with a DC wedge probe (GGB Industries) and controlled by a DAC (Measurement Computing) supplying about 30 mW to each heater. The Si_3N_4 waveguide output is formed as an inverse-taper to edge-couple to a lensed single-mode fibre.

Laser set-up and comb generation procedure. In order to reach mode-locked soliton combs in the first configuration, which uses the dual-cavity design, we first calibrate the laser by aligning the resonances of the two Vernier microring filters using the integrated heaters. This may be done by monitoring the transmitted amplified spontaneous emission (ASE) noise through the filters from the RSOA or by using a separate laser to calibrate the wavelength tuning. Next, the nonlinear microresonator is tuned using its heater to align to the filters' resonances. Once the three are aligned with the pump current above threshold, the device begins to lase. The cavity phase shifter heater, which is positioned over a section of waveguide between the filters and the RSOA, is then tuned to maximize the output power, and the filters may again be adjusted slightly to maximize the output.

After this initialization procedure, the resonance of the nonlinear microresonator is tuned to a longer wavelength such that the original lasing mode is blue-detuned and lasing ceases because the microresonator is no longer on resonance to provide the back-reflection as the laser's output mirror. From this point, the heater is tuned back in the opposite direction to blue-shift the resonance and go through the stages of Fig. 3a–c: first lasing, then chaotic comb generation, and finally soliton states[33]. The resonance producing soliton states corresponds to an effectively red-detuned laser[8], where the detuning results in a typical pump-to-comb conversion efficiency of several per cent[34]. With further tuning of the resonance, output power begins to drop and eventually lasing ceases once the microresonator is fully detuned from the filters and the cavity mode. This procedure allows repeatable generation of soliton states by tuning at rates up to about 10 kHz using a function generator applying a triangle wave voltage to the heater, as shown previously by Joshi et al.[33]; however, we are also often able to reach the soliton states by manual tuning of the heater voltage without a function generator. This relative ease of mode-locking is likely to be a feature of the self-aligning dual-cavity configuration.

In the second, modular configuration, the soliton generation procedure is identical to the first, with the exception that lasing may take place with the nonlinear microresonator off-resonance, allowing a simpler calibration set-up but without the inherent alignment of the microresonator. In both configurations, we were alternatively able to tune the laser from shorter to longer wavelengths across the microresonator resonance using the cavity phase shifter and also achieve soliton mode-locked combs.

Owing to the low pump power needed to generate frequency combs in the microresonator, we did not observe significant thermal shifts in the resonance. If scaled to higher powers where such shifts become stronger, the speed of the resonance tuning can be adjusted to match the power dissipation in the soliton state[8,33].

While we did not require active feedback to maintain the soliton state during our experiments (timescales up to an hour), future systems could account for environmental fluctuations using active feedback[35] to stabilize the soliton states indefinitely. This feedback and the initialization procedure could potentially be controlled using a low-power microcontroller integrated alongside the photonic chip (Fig. 1a) implementing pulse width modulation to efficiently tune the heaters[36].

Comparison of the designs. The two designs demonstrated here, consisting of a dual-cavity comb source and a traditional modular configuration, enable new flexibility in designing the pump laser for generating the frequency comb. The dual-cavity configuration ensures that the microresonator is inherently aligned with the laser because the feedback reflection completes the laser cavity[37]. Detuning the microresonator is still possible, but we observed lower sensitivity to the exact heater settings than found in the modular design, allowing for easier tuning into soliton mode-locked combs through manual tuning (although the automated tuning procedure above was successfully applied to both designs). Additionally, this first design showed a strong output comb power relative to the pump output. Because the microresonator is part of the laser cavity, we cannot directly measure the pump power input to the microresonator, but the theoretical conversion efficiency for solitons[34] suggests that the effective pump input may be notably stronger than the pump output after the microresonator.

The comb generation process in the second, modular design is directly analogous to most previous Kerr comb experiments[8,10,33,38,39]. Despite the potential advantages of the first design, the traditional approach may be desirable if the pump laser and microresonator need to be discretely controlled rather than tuned together. For example, the laser may be locked to a stable reference at a fixed wavelength, simplifying the tuning controls—only the microresonator need be tuned.

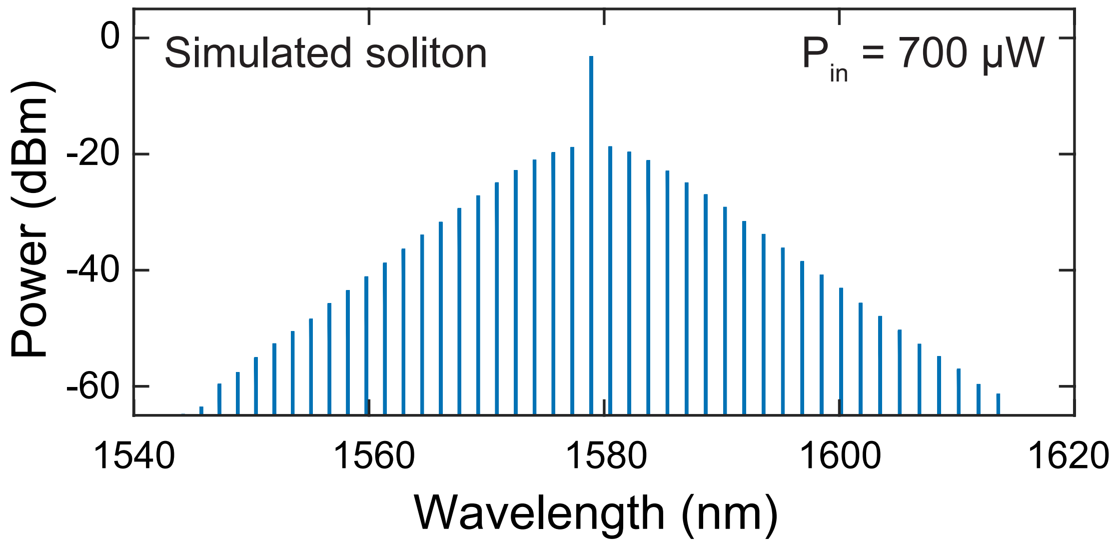
Laser and comb linewidth measurement. The laser linewidth is measured using the delayed self-heterodyne method[28]. The laser output at 80 mA pump current is sent to an interferometer with one path delayed by 12 km of fibre (corresponding to a delay of 58 μs). The other path is phase modulated at 300 MHz. The resulting beat signal is measured on an electrical spectrum analyser (Agilent E4407B) and a 40 kHz Lorentzian linewidth is determined.

The comb linewidth is measured by beating a single comb line with a 1,560 nm 2.4 kHz-linewidth reference laser (Redfern Integrated Optics). With a pump current of 120 mA, a single soliton comb is generated (as in Fig. 3c), and the output is sent to a 50:50 coupler, with the other input coming from the reference laser followed by a polarization controller. The heterodyne output is sent to a photodiode and the RF beat note corresponds to the comb linewidth, which we measure to be approximately 40 kHz, matching that of the pump laser.

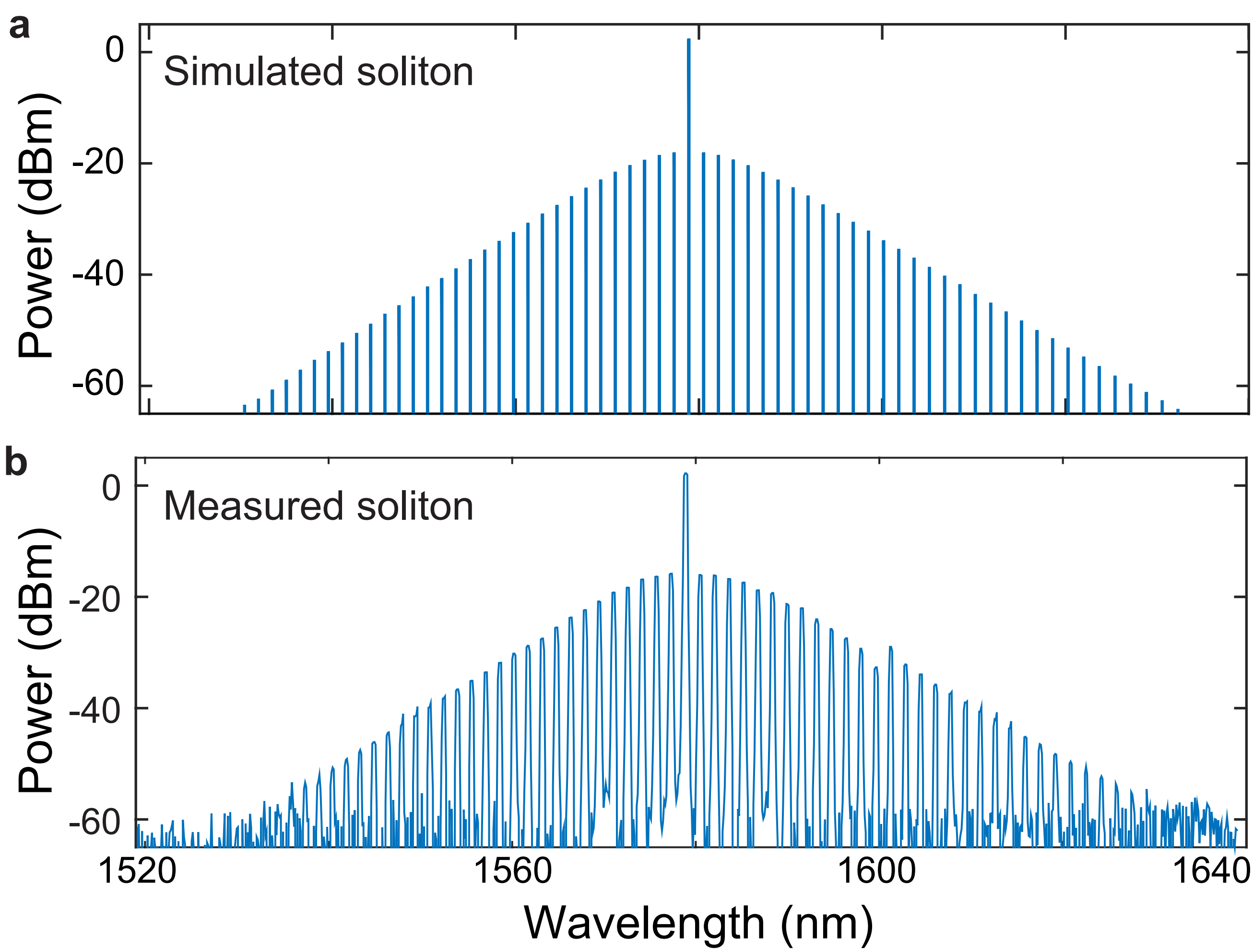
Data availability

The data that support the findings of this study are available from the corresponding authors on reasonable request.

33. Joshi, C. et al. Thermally controlled comb generation and soliton modelocking in microresonators. *Opt. Lett.* **41**, 2565–2568 (2016).
34. Bao, C. et al. Nonlinear conversion efficiency in Kerr frequency comb generation. *Opt. Lett.* **39**, 6126–6129 (2014).
35. Yi, X., Yang, Q.-F., Yang, K. Y. & Vahala, K. Active capture and stabilization of temporal solitons in microresonators. *Opt. Lett.* **41**, 2037–2040 (2016).
36. Cong, G. W. et al. Power-efficient gray-scale control of silicon thermo-optic phase shifters by pulse width modulation using monolithically integrated MOSFET. In *Optical Fiber Communication Conference (2015) M2B.7* (Optical Society of America, 2015).
37. Peccianti, M. et al. Demonstration of a stable ultrafast laser based on a nonlinear microcavity. *Nat. Commun.* **3**, 765 (2012).
38. Hausmann, B. J. M., Bulu, I., Venkataraman, V., Deotare, P. & Lončar, M. Diamond nonlinear photonics. *Nat. Photon.* **8**, 369–374 (2014).
39. Webb, K. E., Erkintalo, M., Coen, S. & Murdoch, S. G. Experimental observation of coherent cavity soliton frequency combs in silica microspheres. *Opt. Lett.* **41**, 4613–4616 (2016).



Extended Data Fig. 1 | Comb generation simulation at low optical power. Shown is the simulated optical spectrum of a soliton comb generated with 700 μW optical pump power (P_{in}) in the bus waveguide before the microresonator. The microresonator dimensions used in the model are 730 nm × 1,800 nm with a radius of 120 μm, corresponding to a 194 GHz FSR.



Extended Data Fig. 2 | Comparison of simulated and measured solitons. **a**, Simulation of a single-soliton comb generated with 2 mW optical pump power in the bus waveguide before the microresonator (1.66 mW after the microresonator). The microresonator dimensions used in the model are 730 nm × 1,800 nm with a radius of 120 μm, corresponding to a 194 GHz FSR. **b**, Optical spectrum of a measured single-soliton comb (from Fig. 3c) with 1.66 mW pump power in the bus waveguide after the microresonator. The sech profile and comb bandwidth qualitatively match those of the simulated comb.

https://doi.org/10.1038/s41586-018-0593-1

Ceramic–metal composites for heat exchangers in concentrated solar power plants

M. Caccia[1,6], M. Tabandeh-Khorshid[1,6], G. Itskos[1,6], A. R. Strayer[1,6], A. S. Caldwell[1], S. Pidaparti[2], S. Singnisai[1], A. D. Rohskopf[2], A. M. Schroeder[3], D. Jarrahbashi[2], T. Kang[2], S. Sahoo[1], N. R. Kadasala[1], A. Marquez-Rossy[4], M. H. Anderson[3], E. Lara-Curzio[4], D. Ranjan[2], A. Henry[2,5] & K. H. Sandhage[1]*

The efficiency of generating electricity from heat using concentrated solar power plants (which use mirrors or lenses to concentrate sunlight in order to drive heat engines, usually involving turbines) may be appreciably increased by operating with higher turbine inlet temperatures, but this would require improved heat exchanger materials. By operating turbines with inlet temperatures above 1,023 kelvin using closed-cycle high-pressure supercritical carbon dioxide (sCO_2) recompression cycles, instead of using conventional (such as subcritical steam Rankine) cycles with inlet temperatures below 823 kelvin[1–3], the relative heat-to-electricity conversion efficiency may be increased by more than 20 per cent. The resulting reduction in the cost of dispatchable electricity from concentrated solar power plants (coupled with thermal energy storage[4–6]) would be an important step towards direct competition with fossil-fuel-based plants and a large reduction in greenhouse gas emissions[7]. However, the inlet temperatures of closed-cycle high-pressure sCO_2 turbine systems are limited[8] by the thermomechanical performance of the compact, metal-alloy-based, printed-circuit-type heat exchangers used to transfer heat to sCO_2. Here we present a robust composite of ceramic (zirconium carbide, ZrC) and the refractory metal tungsten (W) for use in printed-circuit-type heat exchangers at temperatures above 1,023 kelvin[9]. This composite has attractive high-temperature thermal, mechanical and chemical properties and can be processed in a cost-effective manner. We fabricated ZrC/W-based heat exchanger plates with tunable channel patterns by the shape-and-size-preserving chemical conversion of porous tungsten carbide plates. The dense ZrC/W-based composites exhibited failure strengths of over 350 megapascals at 1,073 kelvin, and thermal conductivity values two to three times greater than those of iron- or nickel-based alloys at this temperature. Corrosion resistance to sCO_2 at 1,023 kelvin and 20 megapascals was achieved[10] by bonding a copper layer to the composite surface and adding 50 parts per million carbon monoxide to sCO_2. Techno-economic analyses indicate that ZrC/W-based heat exchangers can strongly outperform nickel-superalloy-based printed-circuit heat exchangers at lower cost.

If concentrated solar power plants with thermal energy storage were to become cost competitive with fossil-fuel plants for electricity generation, then large-scale penetration of renewable solar energy into the electricity grid[11–13] would be enabled, resulting in dramatic reductions in man-made CO_2 emissions (we note that the largest sector-level contributor to global greenhouse-gas emissions is the generation of electricity and heat, with electricity accounting for 68% of this sector[14]). A major technological barrier inhibiting such competitiveness is the development of compact heat exchangers capable of efficient heat transfer at ≥1,023 K for closed-cycle turbine systems operating with high-pressure sCO_2 power cycles. The maximum stresses that can be sustained by the metal alloys used in printed-circuit heat exchangers decline rapidly above 823 K (for example, allowed stresses fall[8] to below 35 MPa at 1,073 K). Heat exchanger materials with enhanced high-temperature failure strength, thermal conductivity, and corrosion resistance that can be cost-effectively manufactured relative to current metal-alloy-based heat exchangers are required.

A key premise of this study is that composites of ceramics and refractory metals can provide a highly attractive combination of properties for robust, cost-effective, compact heat exchangers[9] capable of operating at ≥1,023 K and ≥20 MPa. We demonstrate this here using co-continuous composites of ZrC and W. These are materials with ultrahigh melting points (3,695 K for W and up to 3,700 K for ZrC, respectively[15,16]) that exhibit limited mutual solid solubility, do not react to form other compounds and exhibit a solidus temperature[17] of 3,073 K. Polycrystalline ZrC and W are both thermally conductive[18,19] ($\alpha = 40 \pm 5$ W m^{-1} K^{-1} and 108 ± 13 W m^{-1} K^{-1}, respectively, at 1,000–2,500 K) and exhibit modest values of thermal expansion[20,21] (0.46% and 0.35%, respectively, at 298–1,023 K). As a result, ZrC/W composites are resistant to thermal shock at high heating rates (>1,000 K s^{-1})[22]. While polycrystalline ZrC is stiff and highly creep-resistant[23], polycrystalline W undergoes a brittle-to-ductile transformation[24] at ≤630 K. Consequently, the carbide phase can endow co-continuous ZrC/W composites with high-temperature stiffness, whereas W can provide high-temperature ductility for enhanced resistance to fracture relative to monolithic ZrC.

Co-continuous ZrC/W composites of the desired shapes (bars, disks and plates) were fabricated via the shape-preserving reactive melt infiltration of rigid, porous tungsten carbide (WC) preforms. The WC preforms were prepared by uniaxial compaction of a WC powder/binder mixture followed by heating to 1,673 K and holding for 2 min under an inert (argon; Ar) atmosphere (to allow binder removal and initial-stage sintering or necking of the WC particles). The resulting 52% ± 2% porous WC preforms could readily be machined using standard carbide tooling to generate the desired surface features (channels, headers), as shown in Fig. 1a. The porous WC preforms (Fig. 1b and c) were then heated to 1,373 K in a reducing (4% H_2/Ar) atmosphere, lowered into a molten Zr_2Cu bath held at this temperature (and atmosphere) to allow infiltration, removed from the bulk melt, and then further heated to 1,623 K to promote the following displacement reaction:

$$\frac{1}{2}\mathrm{Zr_2Cu(l)} + \mathrm{WC(s)} \rightarrow \mathrm{ZrC(s)} + \mathrm{W(s)} + \frac{1}{2}\mathrm{Cu(l)} \quad (1)$$

The combined volume of the solid reaction products (ZrC and W) is two times larger than the molar volume of the solid reactant (WC)[25]. Consequently, pores in the rigid WC preforms became filled with the more voluminous solid reaction products, and the non-reactive Cu liquid was forced out, to yield dense ZrC/W-based composites (Fig. 1d–f) that retained the shapes of the starting porous WC preforms

[1]School of Materials Engineering, Purdue University, West Lafayette, IN, USA. [2]George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [3]Department of Engineering Physics, University of Wisconsin, Madison, WI, USA. [4]Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [5]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]These authors contributed equally: M. Caccia, M. Tabandeh-Khorshid, G. Itskos, A. R. Strayer. *e-mail: sandhage@purdue.edu

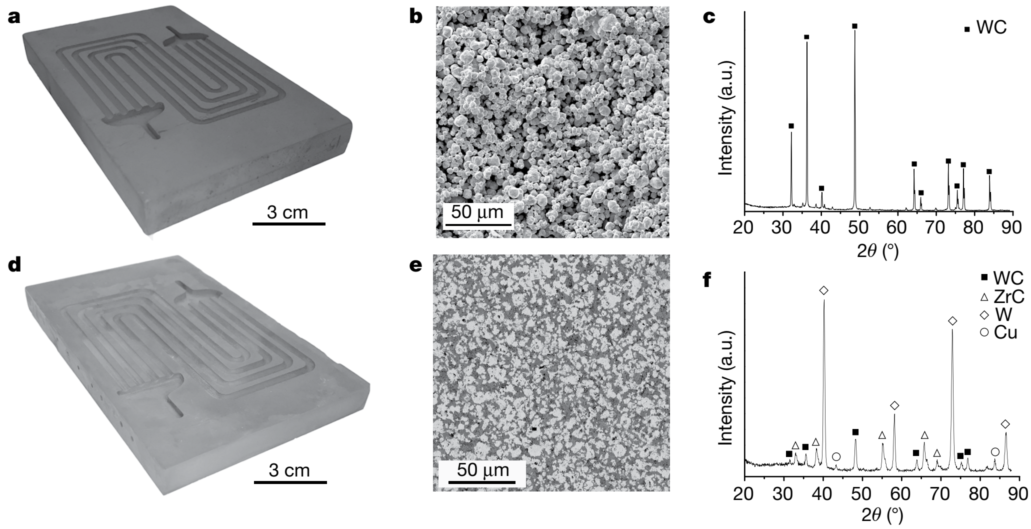


Fig. 1 | Fabrication of dense, channelled ZrC/W plates. a, Photograph of a porous, rigid WC preform plate with four parallel millichannels in a serpentine pattern with two flat-bottom headers. **b**, **c**, Secondary-electron image of a fractured cross-section of a porous, rigid WC preform (**b**) and its corresponding X-ray diffraction pattern (**c**). **d**, Photograph of a dense, channelled ZrC/W-based plate generated by reactive conversion of a rigid, porous, channelled WC plate. **e**, **f**, Backscattered-electron image of a polished cross-section of a dense ZrC/W-based composite prepared by reactive Zr_2Cu(l) infiltration into a porous WC preform (**e**) and its corresponding X-ray diffraction pattern (**f**). a.u., arbitrary units.

(the W particles, along with some unreacted WC, are seen as relatively bright phases, while the ZrC appears as a relatively dark phase in Fig. 1e). Quantitative X-ray diffraction analyses, along with measurements of mass gain upon reactive infiltration, indicated that the ZrC/W-based composites were comprised of 58.1 ± 0.7 vol% ZrC and 35.7 ± 0.4 vol% W, along with residual 2.2 ± 0.6 vol% WC and 4.0 ± 0.7 vol% Cu (average values ± 1 standard deviation; Extended Data Table 1). Comparison of the measured bulk densities ($\rho = 11.15 \pm 0.03$ g cm^{-3}) of these specimens to the theoretical density of this composite ($\rho_{\text{theo}} = 11.43 \pm 0.02$ g cm^{-3}) yielded modest porosity values (<3.0%). Because the reaction-induced increase in solid volume was accommodated by the prior pore volume within the rigid WC preforms, conversion into dense ZrC/W-based composites resulted in average dimensional changes of only $-1.3\% \pm 0.8\%$. The shape- and size-preserving nature of this reactive conversion process allows the cost-effective fabrication of dense, co-continuous ZrC/W composites with well controlled morphologies and dimensions from porous WC preforms that may be readily generated via inexpensive forming operations (which can include stamping and tape casting). Indeed, techno-economic analyses (discussed in more detail in Methods) indicate that such cost-effective processing can allow channelled ZrC/W-based heat exchanger plates to be manufactured at a cost competitive with or lower than that of printed-circuit-type heat exchanger plates comprised of stainless steels or nickel-based superalloys.

The reaction-formed ZrC/W-based composites were found to exhibit attractive thermal and mechanical properties for high-temperature heat exchanger operation. Measurements of the thermal diffusivity[26] ($\alpha = 0.201 \pm 0.013$ cm^2 s^{-1}) and specific heat capacity ($C_p = 0.285 \pm 0.019$ J g^{-1} K^{-1}) of the ZrC/W-based composites yielded an average thermal conductivity ($\kappa = 100\alpha C_p \rho$) of 66.0 ± 4.6 W m^{-1} K^{-1} at 1,073 K, which was 2.5 to 3 times greater than that of Fe-based or Ni-based alloys at this temperature (Extended Data Table 2). The high thermal conductivity of the ZrC/W-based composite relative to Fe-based and Ni-based alloys is of substantial benefit for the performance of compact heat exchangers (that is, higher heat exchanger effectiveness can be achieved for the same heat exchanger geometry). Values of fracture strength were obtained from four-point bend tests[27,28] at room temperature (298 K) and at 1,073 K in an inert (Ar) atmosphere. The bend tests at 1,073 K were conducted both without and after thermal cycling (ten cycles from room temperature to 1,073 K at a rate of 10 K min^{-1}). The failure strengths of the ZrC/W specimens obtained at room temperature, at 1,073 K without thermal cycling, and at 1,073 K after thermal cycling were 348 ± 45 MPa, 369 ± 22 MPa and 387 ± 14MPa, respectively. The comparable values of failure strength obtained at room temperature and at 1,073 K (with and without thermal cycling) were notable given the dramatic reductions in strength above 823 K reported for metal alloys considered for use in heat exchangers for heat transfer to sCO_2 (that is, maximum allowed working stresses of such metal alloys[8] are $\leq$35 MPa at 1,073 K). The retained stiffness and strength of the ZrC/W-based composite at 1,073 K allows the use of thinner heat exchanger plates comprised of this material than plates comprised of Fe- or Ni-based alloys which, in turn, can enhance the high-temperature performance and lower the cost of ZrC/W-based heat exchangers.

Another critical issue is oxidation, given that ZrC and W are not oxidation-resistant at elevated temperatures[29,30]. Here we have developed a strategy for endowing ZrC/W-based composites (or other oxidizable materials[10]) with resistance to high-temperature oxidation by sCO_2-based fluids. Thermodynamic calculations (discussed in more detail in Methods) indicated that the addition of >20 parts per million (p.p.m.) CO to sCO_2 will prevent the oxidation of copper according to the following reaction at $\leq$1,073 K:

$$2\text{Cu(s)} + \{\text{CO}_2\} \leftrightarrow \text{Cu}_2\text{O(s)} + \{\text{CO}\} \quad (2)$$

where $\{CO_2\}$ and {CO} refer to carbon dioxide and carbon monoxide, respectively, present in a fluid solution. Hence, Cu should act as an oxidation-resistant surface layer in such supercritical CO/CO_2 mixtures[10]. CO/CO_2 mixtures have previously been used to lower the effective oxygen partial pressure in hot CO_2-rich gases at ambient pressure, but the addition of CO to sCO_2-based fluids (for buffered supercritical fluids), coupled with the use of a metal surface layer (such as Cu) that is inert in such fluids[10], has not been previously reported, to our knowledge, as a means of suppressing oxidation in such sCO_2-based fluids. ZrC/W-based composite specimens were sealed within Cu by diffusion bonding, as illustrated in Fig. 2a. Corrosion tests were then conducted for up to 1,000 h at 1,023 K and 20 MPa with sCO_2 fluids containing 50 p.p.m. CO. Weight measurements obtained before and after such exposure from five Cu-encapsulated ZrC/W-converted specimens did not yield a detectable weight gain. A backscattered-electron image of a polished specimen cross-section obtained after such exposure, with associated elemental maps for Cu, O, Zr and W, is shown in Fig. 2c. The interface

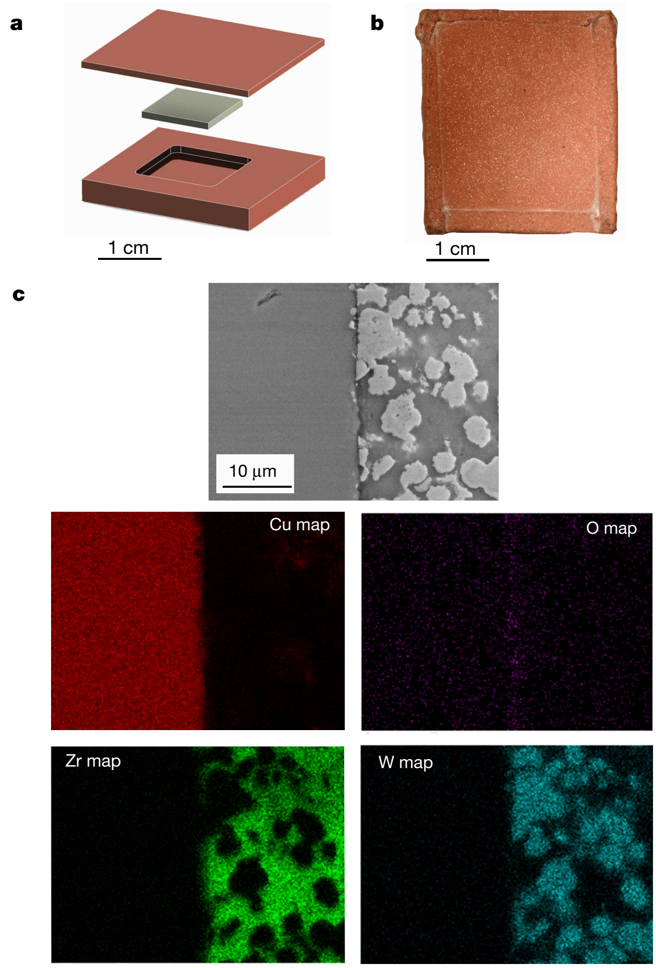


Fig. 2 | Corrosion resistance of a Cu-bearing ZrC/W composite to a buffered supercritical CO/CO_2-bearing fluid at 1,023 K. a, Illustration of the Cu encapsulation of a ZrC/W-converted specimen. **b**, Photograph of a Cu-encapsulated ZrC/W specimen after exposure to a sCO_2 fluid containing 50 p.p.m. CO for 1,000 h at 1,023 K and 20 MPa. **c**, Secondary-electron image and elemental maps for Cu, O, Zr and W of a polished cross-section of such a specimen, after exposure to a sCO_2 fluid containing 50 p.p.m. CO for 1,000 h at 1,023 K and 20 MPa (the top Cu layer was 1 mm thick).

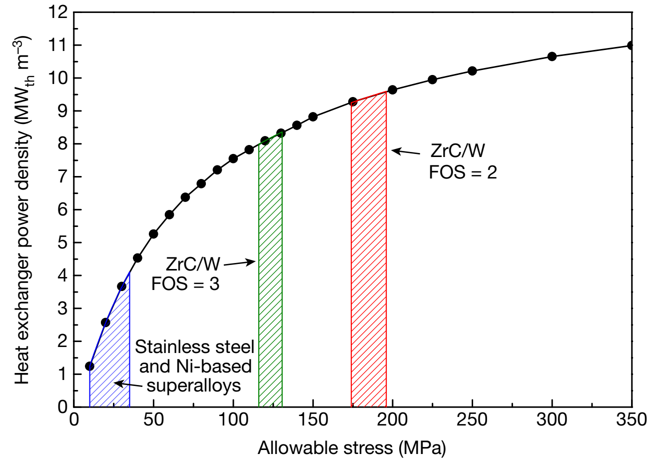


Fig. 3 | Calculated power density of a printed-circuit-type heat exchanger versus allowed stress at 1,073 K. The power density is computed for a 17.5-MW-thermal (MW_{th}) heat exchanger with 95% effectiveness for heat transfer to sCO_2 at 873–1,073 K. As the maximum allowable stress of the material increases, thinner plates may be used, which decreases the required solid volume and increases the power density. Dashed regions correspond to the range of maximum allowed stresses for selected metal alloys[8] (the stainless steels 304 SS and 316 SS and the nickel-based alloys Inconel 617 and 740H) and upper and lower values of failure strengths of ZrC/W-based composites divided by an FOS of 2 or 3.

between the Cu layer and ZrC/W-based composite was free of an apparent continuous Zr-O-based or W-O-based scale, which was consistent with the calculated low oxygen flux through Cu from the CO-bearing sCO_2 fluid at 1,023 K (Methods). These tests confirmed that Cu can act as an effective barrier to inhibit corrosion of ZrC/W-based composites (or other oxidizable materials) in buffered supercritical CO/CO_2 fluids.

The appreciably higher values of thermal conductivity and failure strength of ZrC/W composites relative to Fe- or Ni-based structural alloys at $\geq$1,023 K enable higher values of heat exchanger effectiveness and power density (for the same channel shape, diameter and length). Calculated power density values of a 17.5-MW printed-circuit-type heat exchanger operating with 95% effectiveness for heat transfer to sCO_2 at 873–1,073K via a Brayton cycle are illustrated in Fig. 3 (see Methods). For these calculations, each printed-circuit-type heat exchanger was assumed to contain channels that were straight and parallel, with semicircular cross-sections of diameter 2 mm and lengths of 2.83 m. The thickness of each plate in the printed-circuit-type heat exchanger stack and the spacing between the channels (which yielded the volume of solid material and power density) were then determined from the maximum allowed stresses for each type of material (with a factor of safety, FOS, of 2 or 3 considered for the ZrC/W failure bend strength of 370 MPa). As revealed in Fig. 3, the higher values of the strength for ZrC/W-based composites allow printed-circuit-type heat exchangers to operate at power density values that are at least double those for printed-circuit-type heat exchangers comprised of stainless steels or nickel-based superalloys.

This work demonstrates that cost-effective, reaction-formed composites comprised of co-continuous ceramics and refractory metals, such as ZrC and W, can possess unusual and attractive combinations of properties. The use of such composites in heat exchangers could appreciably enhance the high-temperature performance of renewable concentrated solar power systems, which is an important step towards cost parity with fossil-fuel-derived electricity for reduced greenhouse gas emissions. Although in this paper we have focused on compact ZrC/W-based heat exchangers in concentrated solar power systems operating with high-pressure sCO_2-based power cycles at $\geq$1,023 K, we envisage that cost-effective, reaction-formed ceramic/refractory metal composites could be tailored for use in other desired high-temperature components, enabling more efficient electricity generation in nuclear, natural gas and other power systems.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0593-1.

Received: 11 June; Accepted: 24 August 2018;
Published online 17 October 2018.

1. Dostal, V., Hejzlar, P. & Driscoll, M. J. The supercritical carbon dioxide power cycle: comparison to other advanced power cycles. *Nucl. Technol.* **154**, 283–301 (2006).
2. Weinstein, L. E. et al. Concentrating solar power. *Chem. Rev.* **115**, 12797–12838 (2015).
3. Irwin, L. & Moullec, Y. L. Turbines can use CO_2 to cut CO_2. *Science* **356**, 805–806 (2017).
4. Pitz-Paal, R. Concentrating solar power: still small but learning fast. *Nat. Energy* **2**, 17095 (2017).
5. Lilliestam, J., Labordena, M., Patt, A. & Pfenninger, S. Empirically observed learning rates for concentrating solar power and their responses to regime change. *Nat. Energy* **64**, 17094 (2017).
6. Nithyanandam, K. & Pitchumani, R. Cost and performance analysis of concentrating solar power systems with integrated thermal energy storage. *Energy* **64**, 793–810 (2014).

7. Williams, J. H. et al. The technology path to deep greenhouse gas emissions cuts by 2050. The pivotal role of electricity. *Science* **335**, 53–59 (2012).
8. American Society of Mechanical Engineers (ASME) *2017 ASME Boiler and Pressure Vessel Code. Section II, Part D Properties (Metric)* https://www.asme.org/products/codes-standards/bpvciid-2017-bpvc-section-iimaterialspart (ASME, New York, 2017).
9. Sandhage, K. H. & Henry, A. Methods for manufacturing ceramic and ceramic composite components and components made thereby. International patent application number PCT/US17/28091 (2017).
10. Sandhage, K. H. Method of enhancing corrosion resistance of oxidizable materials and components made therefrom. International patent application number PCT/US17/56015 (2017).
11. Denholm, P. & Maureen, H. Grid flexibility and storage required to achieve very high penetration of variable renewable electricity. *Energy Pol.* **39**, 1817–1830 (2011).
12. Pfenninger, S. et al. Potential for concentrating solar power to provide baseload and dispatchable power. *Nat. Clim. Change* **4**, 689–692 (2014).
13. Denholm, P. & Mehos, M. Enabling greater penetration of solar power via the use of CSP with thermal energy storage. In *Solar Energy: Application, Economics, and Public Perception* (ed. Adaramola, M.) 99–122 (Apple Academic Press, Waretown, 2014).
14. Baumert, K. A., Herzog, T. & Pershing, J. *Navigating the Numbers: Greenhouse Gas Data and International Climate Policy* (World Resources Institute, Washington, 2005).
15. Nagender Naidu, S. V. & Rama Rao, P. *Phase Diagrams of Binary Tungsten Alloys* (ASM International, Materials Park, 1991).
16. Okamoto, H. C-Zr (carbon-zirconium). *J. Phase Equil.* **17**, 162 (1996).
17. Eremenko, V. N., Velikanova, T. Y., Artyukh, L. V., Aksel'rod, G. M. & Vishnevskii, A. S. In *Phase Equilibria Diagrams. Phase Diagrams for Ceramists*. Vol. X. (ed. McHale, A. E.) Fig. 9034 (The American Ceramic Society, Westerville, 1994).
18. Taylor, R. E. Thermal conductivity of zirconium carbide at high temperatures. *J. Am. Ceram. Soc*. **45**, 353–354 (1962).
19. Touloukian, Y. S., Powell, R. W., Ho, C. Y. & Klemens, P. G. in *Thermophysical Properties of Matter* Vol. 1 (Plenum Press, New York, 1970).
20. Touloukian, Y. S., Kirby, R. K., Taylor, R. E. & Lee, T. Y. R. in *Thermophysical Properties of Matter* Vol. 13 (Plenum Press, New York, 1977).
21. Touloukian, Y. S., Kirby, R. K., Taylor, R. E. & Desai, P. D. in *Thermophysical Properties of Matter* Vol. 12 (Plenum Press, New York, 1975).
22. Dickerson, M. B. et al. Near net-shaped, ultra-high melting, recession-resistant ZrC/W-based rocket nozzle liners via the displacive compensation of porosity (DCP) method. *J. Mater. Sci*. **39**, 6005–6015 (2004).
23. Leipold, M. H. & Nielsen, T. H. Mechanical properties of hot-pressed zirconium carbide tested to 2600 °C. *J. Am. Ceram. Soc.* **47**, 419–424 (1964).
24. Lassner, E. & Schubert, W.-D. *Tungsten—Properties, Chemistry, Technology of the Element, Alloys and Chemical Compounds* **22** (Springer, Boston, 1999).
25. Joint Committee on Powder Diffraction Standards (JCPDS) *JCPDS International Center for Diffraction Data File* 00–035–0784 for ZrC, 00–025–1047 for WC, 00–004–0806 for W, http://www.icdd.com/index.php/pdf-4 (JCPDS International Center for Diffraction Data, Newtown Square, 2007).
26. American Society for Testing and Materials (ASTM) *ASTM Standard Test Method for Thermal Diffusivity by the Flash Method* E1461–13 (ASTM International, West Conshohocken, 2013).
27. American Society for Testing and Materials (ASTM) *Standard Test Method for Flexural Strength of Advanced Ceramics at Ambient Temperature* C1161–18 (ASTM International, West Conshohocken, 2018).
28. American Society for Testing and Materials (ASTM) *Standard Test Method for Flexural Strength of Advanced Ceramics at Elevated Temperatures* C1211–18 (ASTM International, West Conshohocken, 2018).
29. Gasparrini, C., Chater, R. J., Horlait, D., Vandeperre, L. & Lee, W. L. Zirconium carbide oxidation: kinetics and oxygen diffusion through the intermediate layer. *J. Am. Ceram. Soc*. **101**, 2638–2652 (2018).
30. Sikka, V. K. & Rosa, C. J. The oxidation kinetics of tungsten and the determination of oxygen diffusion coefficient in tungsten trioxide. *Corros. Sci*. **20**, 1201–1219 (1980).

Acknowledgements This work was supported by the US Department of Energy, Office of Energy Efficiency and Renewable Energy (award number DE-EE0007117). We thank S. H. Hwang for assistance with electron microscopy and I. Itskou for assistance with the preparation of Cu-encased ZrC/W specimens.

Reviewer information *Nature* thanks L. F. Cabeza, O. Graeve, C. Turchi and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.T.-K., A.R.S. and S. Singnisai conducted the WC preform processing and analyses, with assistance from N.R.K. and with K.H.S. providing guidance. M.C. conducted the melt infiltration processing, and analysed the microstructure and chemistry of the resulting ZrC/W composites, with A.R.S., M.T.-K. and A.S.C. providing assistance and with K.H.S. providing guidance. Mechanical tests of ZrC/W composites were conducted and analyzed by S. Sahoo, G.I. and A.M.-R., with E.L.-C. providing guidance. Cu-encased ZrC/W corrosion test specimens were prepared by M.T.-K. Corrosion tests of the Cu-encased ZrC/W specimens were conducted by A.M.S., with M.H.A. providing guidance. Specimen cross-sectional analyses after the corrosion tests were conducted by A.S.C., with assistance from N.R.K. and with K.H.S. providing guidance. Thermodynamic and kinetic calculations associated with corrosion were conducted by K.H.S. Performance calculations of ZrC/W heat exchangers were conducted by S.P., T.K. and D.J., with assistance from K.H.S. and A.H., and with D.R. providing guidance. Economic analyses were conducted by A.D.R., S.P., and A.H., with A.R.S., M.C., G.I. and K.H.S. providing assistance. M.C., G.I., A.H. and K.H.S. drafted the majority of the manuscript. All authors contributed to the writing and review of the manuscript. K.H.S. supervised the overall effort.

Competing interests A.H. and K.S. are inventors on patent applications related to this work that have been filed by (and are owned by) Purdue University and the Georgia Institute of Technology (see refs [10,11]). Patent application number PCT/US17/28091 includes the fabrication and use of ZrC/W composites for heat exchangers. Patent application PCT/US17/56015 includes enhancement of the high-temperature oxidation resistance of ZrC/W composites through the use of carbon monoxide-bearing supercritical carbon dioxide and a copper surface layer. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0593-1.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.H.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Fabrication of rigid, porous WC preforms. Shaped green bodies were generated from mixtures of WC powder (5–6-μm-diameter particle size; SC55S, Global Tungsten & Powders) and isobutyl methacrylate (IBMA, Elvacite 2045, Lucite International) as a binder. A slurry comprised of 89 wt% WC powder and an 11 wt% IBMA/acetone solution (11 wt% IBMA dissolved in acetone) was first prepared. The slurry was placed, along with WC milling media (9.5-mm-diameter balls; WC ball:slurry weight ratio of 4.4:1), in the tank of a water-cooled attrition mill (SC-1 attritor, Union Process) and then mixed for 10 min with the WC impeller of the attrition mill rotating at 80 r.p.m. After drying for 12 h at room temperature, the resulting solid WC/IBMA powder mixture was milled (ball:powder weight ratio of 4.9:1) for an additional 30 s at 120 r.p.m. (to break up agglomerates) and then passed through a 200-mesh sieve. The WC/IBMA powder mixture (1.4 wt% IBMA) was then formed into plates (9 cm × 15 cm × 1 cm) or disks (12.5 mm diameter × 3 mm thickness; 4.85 mm diameter × 3.8 mm thickness) via compaction within single-action, graphite-lubricated steel dies of appropriate shape at a peak pressure of 10 MPa at room temperature. The compacted WC/IBMA specimens were then heated in a flowing high-purity Ar atmosphere at 2.5 K min^{-1} up to a peak temperature of 1,673 K (to allow binder removal) and held at this temperature for 2 h (to allow initial-stage sintering or necking of the WC particles) to yield rigid WC preforms with bulk porosity values of 52% ± 2%. Green machining of the surfaces of rigid, porous WC preforms was conducted with a computer numerical-controlled (CNC) milling machine (Miyano TSV-35, Precision Technologies) at a traverse rate of 38 cm min^{-1} using a 3-mm carbide ball tool rotating at 5,000 r.p.m. (to generate channels of semi-circular cross-section) and a flat-bottomed carbide tool (to generate flat-bottomed headers).

Conversion of porous WC preforms into dense ZrC/W composites. The rigid, porous WC preforms were transformed into dense ZrC/W composites via pressureless infiltration and reaction with liquid Zr_2Cu. Alloying of the Zr_2Cu liquid and the reactive melt infiltration process were conducted within a controlled-atmosphere, electrically heated, vertical tube furnace (mullite tube). The Zr_2Cu liquid was prepared by melting stoichiometric mixtures of Zr (≥99.2% purity; Zirconium Research Corp.) and Cu (≥99.9% purity; McMaster Carr) at 1,473 K for ≥2 h in a flowing 4% H_2/Ar atmosphere within a graphite crucible. To allow immersion and retrieval from the Zr_2Cu liquid bath, the porous WC preform plates or disks were lowered and raised in a vertical orientation within an open graphite support frame consisting of two horizontal plates (a top plate and a bottom plate) and two vertical (side wall) plates. The top horizontal graphite plate contained a hole for connection to a vertical graphite ram (for lowering and raising the graphite frame containing the porous WC preform into the Zr_2Cu bath). After heating the WC preform and Zr_2Cu(l) to 1,373 K (above the 1,273 K congruent melting temperature of Zr_2Cu) in a flowing 4% H_2/Ar atmosphere, the support frame was lowered at a rate of 5 cm min^{-1} into the Zr_2Cu melt bath so as to completely immerse the porous WC preform. After immersion, the infiltrated preform was raised out of the Zr_2Cu melt bath at a rate of 5 cm min^{-1} and then heated at 1 K min^{-1} to 1,623 K and held at this temperature for 2 h to further the conversion reaction of WC into the ZrC/W composite. After cooling to room temperature within the flowing 4% H_2/Ar atmosphere, the reacted specimen was removed from the vertical tube furnace.

Macro- and micro-structural characterization of the ZrC/W composites. Archimedes measurements, using water as the buoyant fluid, were used to obtain density values for the ZrC/W-converted specimens. The overall dimensions and the dimensions of surface features (channels, headers), of the converted specimens were obtained with the use of a digital caliper (CD-6″ CSX ABSOLUTE Digimatic, Mitutoyo) to allow comparison with the corresponding dimensions of the starting porous WC preform. The ZrC/W-converted specimens were cross-sectioned by a diamond sectioning saw (Techcut 4 Precision Low Speed Saw, Allied High Tech Products Inc.) or by electrodischarge machining (FX20K CNC Wire EDM, Mitsubishi Electric Corp.). Such cross-sections were mounted in epoxy and then ground and polished with a series of diamond pastes to a surface finish of 1 μm. Microstructural analyses of these polished cross-sections were conducted with a field-emission-gun scanning electron microscope (S-4800 FE-SEM, Hitachi) equipped with an energy dispersive X-ray spectrometer (INCA EDS, Oxford Instruments). X-ray diffraction analyses of the phase contents of the ZrC/W-converted specimens were conducted with Cu Kα radiation at a scan rate of 3 degrees per minute (D2 Phaser diffractometer, Bruker). To allow quantification of the phase content, X-ray diffraction calibration curves were generated from mixtures of W, WC and ZrC powders combined in proportions consistent with the stoichiometry of reaction (1) for various degrees of reaction.

Quantitative phase analyses of ZrC/W-converted specimens. The reaction of Zr_2Cu liquid within an infiltrated WC preform to form ZrC and W, with residual retained Cu and WC, can be described by the following reaction:

$$\frac{a}{2}\mathrm{Zr_2Cu(l)}+\mathrm{WC(s)} \rightarrow a\mathrm{ZrC(s)} + a\mathrm{W(s)} + (1-a)\mathrm{WC(s)} + b\mathrm{Cu(s)} + \left(\frac{a}{2}-b\right)\mathrm{Cu(l)} \quad (3)$$

where Cu(s) refers to solid copper retained in the specimen, and Cu(l) refers to liquid copper extruded back out of the specimen (upon filling of the prior pores with the solid ZrC and W products of this reaction). The value of the parameter a for this reaction (an indication of the extent of reaction) is related to the molar W:WC ratio in a reacted specimen as follows:

$$\frac{\mathrm{W}}{\mathrm{WC}} = \frac{a}{(1-a)}$$

The molar W:WC ratio was determined from quantitative X-ray diffraction analysis of the reacted specimen; that is, the ratio of the areas A under the (110) and (100) X-ray diffraction peaks for W and WC, respectively, was determined and then a calibration curve (providing the correlation between this $A_{110(\mathrm{W})}/A_{100(\mathrm{WC})}$ ratio and the molar W:WC ratio[31]) was used to determine the molar W:WC ratio. The value of the parameter b was then obtained from the measured mass gain, Δm, of the infiltrated/reacted specimen, according to the following equation:

$$\frac{\Delta m}{m_o} = \frac{[a\mathrm{MW_{ZrC}} + a\mathrm{AW_W} + (1-a)\mathrm{MW_{WC}} + b\mathrm{AW_{Cu}}] - \mathrm{MW_{WC}}}{\mathrm{MW_{WC}}}$$

where m_o refers to the starting mass of the porous (non-infiltrated) WC preform, MW_i refers to the molecular weight of compound i, and AW_j refers to the atomic weight of element j. Parameters a and b, along with the molar volumes[25,32] of ZrC, WC, W and Cu, were used to determine values of the volume percentages of these phases in reacted specimens. The theoretical densities of non-porous composites comprised of such ZrC, W, WC and Cu mixtures were then compared to measured specimen densities to obtain bulk porosity values. The experimentally determined values of the extent of reaction, phase content, density and porosity for five reactively converted specimens are provided in Extended Data Table 1.

Evaluation of the thermal conductivity of ZrC/W composites. A ZrC/W-converted disk was cut by electrodischarge machining into cylindrical specimens with: (1) 12.5 mm diameter and 3 mm thickness (for thermal diffusivity analyses) and (2) 4.85 mm diameter and 3.8 mm thickness (for heat capacity analyses). Laser flash measurements[26] (Flashline 4010 Thermal Properties Analyzer, Anter Corporation) were used to obtain thermal diffusivity values, and differential scanning calorimetry (LABSYS Evo TG-DTA-DSC Analyzer, Setaram Instrumentation) was used to evaluate the heat capacity of ZrC/W-converted specimens at 1,073 K (both analyses were conducted at the Orton Materials Testing and Research Center, Westerville, Ohio, USA). These measurements then yielded the value of thermal conductivity at 1,073 K according to the equation $\kappa = 100\alpha C_p \rho$, with κ the thermal conductivity (in units of W m^{-1} K^{-1}), α the thermal diffusivity (cm^2 s^{-1}), ρ the density (g cm^{-3}), and C_p the specific heat capacity (J g^{-1} K^{-1}). The average thermal conductivity and the 95% confidence limit range were determined from measurements conducted on nine separate specimens.

Evaluation of the fracture strength of ZrC/W composites. A ZrC/W-converted disk was cut by electrodischarge machining into bars of rectangular cross-section, and then ground and chamfered (Bomas Machine Specialties Inc.) to yield four-point bend test specimens with dimensions[27,28] of 2 mm (±0.05 mm) × 1.5 mm (±0.05 mm) × 25 mm. Fracture strength tests were conducted at 298 K, at 1,073 K and at 1,073 K after thermal cycling. The latter specimens were cycled ten times between 298 K and 1,073 K at a heating rate of 10 K min^{-1} in a high-purity Ar atmosphere. Each test specimen was loaded onto a four-point test fixture (loading span of 20 mm, support span of 10 mm [27,28]) contained in a controlled atmosphere furnace. For the tests conducted at 1,073 K, the furnace chamber was evacuated and backfilled three times with 4% H_2/Ar gas. The temperature was then raised to 1,073 K at a rate of 10 K min^{-1} under flowing 4% H_2/Ar gas. An increasing force was applied with a crosshead speed of 0.0083 mm s^{-1}, with the force measured continuously to the point of specimen failure. The fracture strength was calculated using the following equation:

$$\sigma_F = \frac{3PL}{4bd^2}$$

with P the break force, L the outer span, b the specimen width and d the specimen thickness. For each of the three test conditions (298 K, 1,073 K and 1,073 K after thermal cycling), the average fracture strength value, and the 95% confidence

limit range, were determined from measurements conducted on ten separate specimens.

Thermal and mechanical properties of ZrC/W, IN740H and IN617. The solidus temperature, as well as the high-temperature (1,073 K) values of the specific heat capacity, the thermal conductivity and the failure strength, for ZrC/W and for the high-temperature nickel-based superalloys, Inconel 740H and Inconel 617, are provided in Extended Data Table 2.

Corrosion resistance of Cu-covered ZrC/W in CO-bearing sCO_2. Corrosion testing of Cu-encased ZrC/W specimens was conducted in a supercritical CO-bearing CO_2 mixture at 1,023 K and 20 MPa. Copper encapsulation was conducted by placing a 10 mm × 10 mm × 1 mm ZrC/W-converted specimen inside a cavity machined into a copper plate (Fig. 2a). A sheet of copper (0.5 mm thick or 1 mm thick) was placed on top of the ZrC/W sample and the copper plate, and hot pressing was then conducted at 1,193 K at 10 MPa for 2 h to bond the copper sheet to the sample and the copper plate. Five such Cu-encapsulated ZrC/W specimens were then exposed to a flowing mixture (around 0.10 $kg\,h^{-1}$) of CO_2 with 50 ± 10 p.p.m. CO (as measured with a gas chromatograph) at 1,023 K and 20 MPa for 1,000 h. Weight change measurements were obtained before and after such exposure for each specimen. The specimens were also cross-sectioned by a diamond sectioning saw, mounted in epoxy, and then polished to a 1-μm finish to allow analyses with a field-emission-gun scanning electron microscope.

Analysis of Cu compatibility with CO-bearing sCO_2 mixtures at ≤1,073 K. Thermodynamic calculations indicated that small additions of CO to the sCO_2 could reduce the oxygen fugacity to values sufficiently low to avoid the oxidation of certain metals, such as copper. Consider the following reaction:

$$2\mathrm{Cu(s)} + \{\mathrm{CO_2}\} \leftrightarrow \mathrm{Cu_2O(s)} + \{\mathrm{CO}\} \quad (4)$$

where $\{CO_2\}$ and $\{CO\}$ refer to carbon dioxide and carbon monoxide, respectively, present in a fluid solution. For pure Cu and pure stoichiometric Cu_2O present in their pure reference states, the equilibrium CO/CO_2 fugacity ratio for this reaction at 1,073 K and 1 atm total pressure is 15.3×10^{-6} (at 1,023 K, this ratio decreases to 8.12×10^{-6})[33]. With equal moles of reactant fluid (CO_2) and product fluid (CO) on both sides of this reaction, the volume V change per mole of reaction should be relatively small and may be positive[32] (since the volume of one mole of Cu_2O is larger than the volume of 2 moles of Cu). Consequently, the change in the Gibbs free energy for this reaction with total pressure P at a given temperature T ($\partial\Delta G_{rxn}/\partial P|_T = \Delta V_{rxn}$) should be relatively small and probably positive (that is, the equilibrium for this reaction as written above is likely to shift slightly to the left with increasing total pressure), so that the equilibrium CO/CO_2 fugacity ratio for this reaction should probably decrease slightly as the total pressure increases. Hence, with the addition of 20 p.p.m. or more of CO to sCO_2, we expected that the oxidation of Cu in such a CO-bearing sCO_2-based fluid at ≤1,073 K and 20 MPa would be avoided[10]. This expectation was confirmed by the lack of a detectable weight gain (due to oxidation) of a Cu-encapsulated specimen after exposure to a mixture of 50 p.p.m. CO in sCO_2 for 1,000 h at 1,023 K and 20 MPa.

Analysis of the oxygen flux through Cu. Although the thermodynamic analysis above indicates that Cu should not oxidize in sCO_2 containing 50 p.p.m. CO at ≤1,073 K, the low concentration of oxygen present in such a fluid at equilibrium may dissolve into, and migrate through, solid Cu. Hence, a calculation of the maximum oxygen flux through Cu exposed to such a mixture of CO and CO_2 was conducted.

The diffusivity of oxygen D_O (in units of $cm^2\,s^{-1}$) in copper (for temperatures in the range of 973 K to 1,573 K) has been reported to obey the equation[34]:

$$D_O = 1.16 \times 10^{-2} \times \exp(-67,300JR^{-1}T^{-1})$$

The solubility of oxygen X_O (the oxygen atom fraction) in copper (for the oxygen partial pressure established by the Cu/Cu_2O equilibrium) has been reported to obey the equation[34]:

$$X_O = 154 \times \exp(-149,600JR^{-1}T^{-1})$$

These equations yield, at 1,023 K, $D_O = 4.25 \times 10^{-6}\,cm^2\,s^{-1}$ and $X_O = 3.54 \times 10^{-6}$.

The standard Gibbs free energy changes per mole of the following reactions:

$$\{\mathrm{CO}\} + \frac{1}{2}\{\mathrm{O_2}\} \leftrightarrow \{\mathrm{CO_2}\} \quad (5)$$

$$2\mathrm{Cu(s)} + \frac{1}{2}\{\mathrm{O_2}\} \leftrightarrow \mathrm{Cu_2O(s)} \quad (6)$$

at 1,023 K and 1 atm are $\Delta G^o_{rxn(5)} = -193{,}551$ J and $\Delta G^o_{rxn(6)} = -93{,}854$ J (ref. [33]). Hence, the equilibrium oxygen fugacity associated with reaction (5) at a f_{CO2}/f_{CO} ratio of 0.99995/0.00005 (50 p.p.m. CO in CO_2) at 1,023 K and 1 atm total pressure is 6.87×10^{-12} atm. The equilibrium oxygen fugacity associated with reaction (6) at 1,023 K and 1 atm total pressure is 2.60×10^{-10} atm. According to Sievert's law[35] (which applies for low oxygen contents where Henry's law is valid, as should be the case here for reactions (5) and (6) at 1,023 K), the solubility of a diatomic gas in a condensed phase should vary with the square root of the gas partial pressure (or fugacity). Hence the solubility of oxygen in copper that is equilibrated with a 50 p.p.m. CO-bearing sCO_2 mixture at 1,023 K should be (ignoring non-ideal behaviour for the CO-bearing sCO_2 mixture, as an approximation):

$$\frac{X_O(\text{at fraction for 50 p.p.m. CO/sCO}_2\text{ equilibrium})}{X_O(\text{at fraction for Cu/Cu}_2\text{O equilibrium})} = [(6.87 \times 10^{-12})/(2.60 \times 10^{-10})]^{1/2}$$

or

$$X_O(\text{at fraction for 50 p.p.m. CO/sCO}_2\text{ equilibrium})/3.54 \times 10^{-6} = [(6.87 \times 10^{-12})/(2.60 \times 10^{-10})]^{1/2}$$

or

$$X_O(\text{at fraction for 50 p.p.m. CO/sCO}_2\text{ equilibrium}) = 5.75 \times 10^{-7}$$

Consider a dense layer of Cu on top of a ZrC/W composite exposed to such a 50 p.p.m. CO-bearing sCO_2 mixture at 1,023 K. If a linear concentration gradient is assumed for oxygen through the copper layer under steady-state conditions (that is, assuming that the diffusion of oxygen in copper is independent of the oxygen concentration and that chemical reactions at both the CO/sCO_2:Cu and Cu:ZrC/W interfaces are at local equilibrium), then the approximate atomic flux of oxygen J_O (in units of moles of oxygen $cm^{-2}\,s^{-1}$) through such a copper layer may be expressed as:

$$J_O = -D_O\Delta X_o/[LV_m(\mathrm{Cu})]$$

where ΔX_o is the difference between the atomic fraction of oxygen dissolved in copper at the Cu:ZrC/W interface and that at the Cu:CO/sCO_2 interface; L is the thickness of the copper layer; and $V_m(Cu)$ is the molar volume (in units of cm^3 per mole) of copper. The maximum flux of oxygen would occur if it is assumed that the mole fraction of oxygen dissolved in Cu at the Cu:ZrC/W interface is zero. Hence, the maximum oxygen flux is given by:

$$J_O^{max} = D_O X_o/[LV_m(\mathrm{Cu})]$$

Using a Cu molar volume[32] of 7.11 cm^3 per mole, along with the values of D_O and X_o calculated above, the maximum steady-state flux of oxygen through a Cu layer of 1,000 μm (0.1 cm) thickness at 1,023 K in a 50 p.p.m. CO-bearing sCO_2 mixture is thus:

$$J_O^{max} = 3.44 \times 10^{-12}\ \text{moles O cm}^{-2}\,\text{s}^{-1}$$

For this value of the oxygen flux, 124×10^{-5} moles of O per cm^2 (or effectively 6.20×10^{-6} moles of O_2 per cm^2) would migrate through such a Cu layer in 1,000 h (3.6×10^6 s). If it is assumed that all of the oxygen arriving at the Cu:ZrC/W interface reacted with ZrC/W to generate either monoclinic ZrO_2(s) or WO_2(s) products by the following reactions, then the corresponding thicknesses of ZrO_2 or WO_2 layers (for ZrO_2 and WO_2 molar volumes of 21.2 cm^3 per mole and 19.6 cm^3 per mole, respectively[32]) would be only 1.3 μm or 1.2 μm, respectively.

$$\mathrm{ZrC(s)} + 2\{\mathrm{O}\} \leftrightarrow \mathrm{ZrO_2(s)} + \mathrm{C(s)} \quad (7)$$

$$\mathrm{W(s)} + 2\{\mathrm{O}\} \leftrightarrow \mathrm{WO_2(s)} \quad (8)$$

Given the assumptions made in this maximum oxygen flux calculation (for example, rapid chemical reactions and local equilibrium at the CO/sCO_2:Cu and Cu:ZrC/W interfaces, complete consumption of all of the oxygen arriving at the Cu:ZrC/W to form ZrO_2 or WO_2, composition-independent diffusion coefficient for oxygen transport through Cu), the modest amount of oxide formation predicted was consistent with the negligible amount of oxide detected at the Cu:ZrC/W interface (Fig. 2) after exposure for 1,000 h to a mixture of 50 p.p.m. CO in sCO_2 at 1,023 K and 20 MPa.

Performance analyses of ZrC/W-based heat exchangers. Performance calculations were conducted for a 17.5-MW-thermal (MW_{th}) counter-flow compact (printed-circuit-type) heat exchanger, designed to be representative of an intermediate heat exchanger for a 10-MW-electric (MW_e) concentrated solar power plant. The simulated concentrated solar power plant used a molten KCl–$MgCl_2$ salt[36,37] as the thermal energy storage medium and a sCO_2 Brayton cycle for power generation, as depicted in Extended Data Fig. 1.

Thermal-hydraulic design simulations. The operating conditions for the molten salt and sCO_2 in the $Q = 17.5$ MW_{th} intermediate heat exchanger are provided in Extended Data Table 3. These operating conditions corresponded to a log mean temperature difference (LMTD) of 10 K, for a desired UA value of 1,750 kW K^{-1} (as in the following equation), and a heat exchanger effectiveness of 95%.

$$Q = UA \times LMTD$$

For the present calculations, it was assumed that the channels in the compact heat exchanger were straight and had a semi-circular cross-section with a diameter (maximum channel width) of 2 mm. The heat exchanger plate width was fixed at 0.60 m. The Reynolds numbers, Re, for the flows of molten salt and sCO_2 were calculated with the assumption that the heat exchanger possessed the same number of channels for both fluids. Published correlations for such compact heat exchangers[38] were then used to calculate the overall heat exchanger heat transfer coefficient (U). From the values of UA and U, the total heat transfer area (A) of the heat exchanger was determined. For semi-circular channels of fixed diameter (2 mm), the required heat exchanger area (A) could be achieved with different combinations of the channel length and the number of channels. The values of pressure drop for each fluid, for the operating conditions shown in Extended Data Table 3, were then calculated for various combinations of channel length and channel number[38]. As an example, the influence of the channel number (for a fixed UA value of 1750 kW K^{-1}) on the pressure drop for each fluid is shown in Extended Data Fig. 2.

Mechanical design calculations. Once the required values of heat transfer coefficient (U), heat exchanger area (A), and associated combinations of channel length and channel number were determined from the thermal-hydraulic design calculations, the total volume of solid material (non-channel volume) in the compact heat exchanger was determined by the pressure containment requirements between the hot and cold streams. The mechanical design calculations were performed to comply with the ASME Boiler and Pressure Vessel Code Section VIII standard for diffusion bonded heat exchangers; that is, the minimum values of the plate thickness and the channel spacing required to stay within certain values of membrane and bending stresses were calculated (Extended Data Fig. 3)[39]. The values of plate thickness and channel spacing for given allowed stress values (for heat exchangers with straight channels of semi-circular cross-section and a fixed diameter of 2 mm) were then used, along with the desired UA and A values, to determine the total volume (combined solid and channel volumes) of the heat exchanger (in m^3) and the associated value of heat exchanger power density (in $MW_{th}\,m^{-3}$). For a variety of stainless steels (such as 304 SS and 316 SS) and nickel-based alloys (such as Inconel 617 and 740H), the maximum allowed stresses at 1,073 K are reported to be in the range[8] 10–35 MPa. The measured failure strengths of the ZrC/W composites of the present work at 1,073 K were 369 ± 22 MPa (with ± 22 MPa referring to the 95% confidence interval range). Assuming an FOS value of 2 to 3 for these ceramic/metal composites (that is, failure strengths of 123 ± 7 MPa and 185 ± 11 MPa, respectively), the calculated values of power density of the ZrC/W-based heat exchangers were found to be a factor of two or more greater than for the metal alloys (Fig. 3) for the operating conditions indicated in Extended Data Table 3.

For example, a power density of 8.2 $MW_{th}\,m^{-3}$, and pressure drop values of 100 kPa for sCO_2 and 39 kPa for the molten KCl-$MgCl_2$ salt, could be achieved with a ZrC/W-based heat exchanger (FOS of 3) with overall dimensions (width × length × height) of 0.60 m × 2.83 m × 1.25 m (comprised of a stack of 804 plates, each of 1.52 mm thickness, with semi-circular channels of 2 mm diameter and centre-to-centre spacing of 2.51 mm, and 182,400 total channels or 91,200 channels for each fluid). For a heat exchanger comprised of Inconel 740H (with a maximum allowed stress[8,40] of 35 MPa at 1,073 K), such calculations yielded a power density up to 4.1 $MW_{th}\,m^{-3}$, with similar pressure drop values, for overall dimensions (width × length × height) of 0.60 m × 2.83 m × 2.49 m (comprised of a stack of 1,215 plates, each of 2.03 mm thickness with semi-circular channels of 2 mm diameter and centre-to-centre spacing of 3.80 mm and 91,200 channels for each fluid). For a heat exchanger comprised of 316 SS (at a maximum allowed stress of 10 MPa at 1,073K)[8], such calculations yielded a power density of only up to 1.2 $MW_{th}\,m^{-3}$, with similar pressure drop values, for overall dimensions (width × length × height) of 0.60 m × 2.83 m × 8.32 m (comprised of a stack of 2,656 plates, each of 3.12 mm thickness with semi-circular channels of 2 mm diameter and centre-to-centre spacing of 8.30 mm and 91,200 channels for each fluid).

Techno-economic analyses of ZrC/W-based heat exchangers. ZrC/W-based composites were fabricated via the reactive infiltration of WC powder-derived preforms with Zr_2Cu liquid. Direct supplier quotes for bulk quantities of WC, Zr and Cu yielded costs (costs throughout are given in 2018 US dollars) of \$21.7 kg^{-1}, \$23.6 kg^{-1}, and \$6.9 kg^{-1}, respectively, for these raw materials. The combined cost of these starting materials consumed to produce the ZrC/W-based composites was \$21.8 kg^{-1}, which equates to \$249,000 m^{-3} (for a composite theoretical density of 11,430 kg m^{-3}; Extended Data Table 1). The cost of Cu foil (used to coat the ZrC/W composites for resistance to corrosion; Fig. 2) was \$10 kg^{-1} (supplier quotes). Hence, the cost of these starting raw materials consumed in the production of a 17.5-MW_{th} counter-flow compact (printed-circuit-type) ZrC/W-based heat exchanger with overall dimensions of 0.60 m × 2.83 m × 1.25 m (width × length × height) was \$0.037 W^{-1} (for the heat exchanger discussed above, comprised of a stack of 804 plates, each of 1.52 mm thickness, with semi-circular channels of 2 mm diameter and centre-to-centre spacing of 2.51 mm, and with Cu foil used to coat the channels for the sCO_2-based fluid).

An Excel-based model developed by Ricardo Strategic Consulting for the US Department of Energy was used to evaluate the processing costs (that is, costs associated with producing the final product other than the aforementioned raw material costs). This model utilizes a range of inputs, including the required manufacturing equipment, installation costs, the space and energy needed to operate equipment, the level and associated cost of skilled labour, the time required to process each component, the mean time to failure or replacement of equipment and maintenance costs, indirect costs and other parameters. A conceptual layout was developed for a facility containing the required equipment (for example, for powder/binder mixing, powder compaction, binder removal, preform sintering, preform machining, melt infiltration, diffusion bonding and brazing/welding equipment) necessary to manufacture 100 heat exchanger plates per day (for plates with dimensions of 0.60 m width × 2.5–3 m length × 1–2 mm thickness). After inserting the costs of equipment, labour, energy, maintenance and other inputs into the Ricardo Excel model, it was apparent that the capital equipment expenditure was the dominant processing cost factor. Since the relative contribution of the capital equipment expenditure to the total cost (raw materials + processing) of a given heat exchanger will depend on the rate of production of the heat exchangers (and, in turn, the rate of production of heat exchanger plates), the Ricardo excel model was used to evaluate the heat exchanger plate production rate required for the processing cost to fall well below the cost of raw materials. These calculations indicated that, at a production rate of 10,000 heat exchanger plates per year, the processing cost would fall below 15% of the cost of the raw materials; that is, for a production rate of >10,000 plates per year, the total cost of a counter-flow compact (printed-circuit-type) ZrC/W-based heat exchanger would be <\$0.043 W^{-1} (<\$25.1 kg^{-1}).

This estimated manufacturing cost for such ZrC/W-based heat exchangers compared favourably to the cost of Ni superalloy-based compact heat exchangers. For example, a compact (printed-circuit-type) heat exchanger comprised of Inconel 740H (IN740H) was considered. The maximum allowed stress[8,40] of IN740H at 1,073 K is 35 MPa, which allows the use of a 17.5-MW_{th} counter-flow compact heat exchanger with a power density of up to 4.1 $MW_{th}\,m^{-3}$ (a relatively high value for metal-alloy-based heat exchangers, Fig. 3). This Ni-based alloy contains[40] 24.5 wt% Cr, 20 wt% Co, 1.5 wt% Nb, 1.35 wt% Al, 1.35 wt% Ti, 0.15 wt% Si, 0.1 wt% Mo and 0.03 wt% C. The commodity prices (http://www.infomine.com/) of raw materials for the three major constituents of this alloy (\$14.6 kg^{-1} for nickel, \$4.7 kg^{-1} for chromium from ferrochrome at \$2.8 kg^{-1}, \$91.0 kg^{-1} for cobalt) yielded a combined raw materials commodity cost of at least \$26.8 kg^{-1} for IN740H. The solid volume of IN740H in the 17.5 MW_{th} heat exchanger discussed above (with a power density up to 4.1 $MW_{th}\,m^{-3}$) with overall dimensions (width × length × height) of 0.60 m × 2.83 m × 2.49 m (comprised of a stack of 1,215 plates, each of 2.03 mm thickness with 91,200 semi-circular channels of 2 mm diameter for each fluid) would be 3.42 m^3, that is: $[0.6\text{ m} \times 2.83\text{ m} \times 2.49\text{ m}] - [\pi(0.001\text{ m})^2 \times 2.83\text{ m} \times 0.5 \times 91{,}200 \times 2]$. For an IN740H density[40] of 8,050 kg m^{-3}, such a 17.5-MW_{th} IN740H heat exchanger would have mass 2.75×10^4 kg. Hence, the raw materials commodity cost alone of such a 17.5-MW_{th} IN740H heat exchanger would be at least \$0.042 W^{-1} (\$26.8 $kg^{-1} \times 2.75 \times 10^4$ kg/17.5×10^6 W). A similar calculation for the four major constituents of IN617 (44.5 wt% Ni, 22 wt% Cr, 12.5 wt% Co, 9 wt% Mo, with smaller amounts of Fe, Al, Mn, Si, Ti, Cu, C, and B), for which the maximum allowed stress[8,41] at 1,073 K is 32 MPa, yielded a raw materials commodity cost of at least \$22.4 kg^{-1} for this alloy (with \$39 kg^{-1} for molybdenum from molybdenum trioxide at a commodity cost of \$26.0 kg^{-1}; http://www.infomine.com/), which corresponded to a cost of at least \$0.037 W^{-1} (\$22.4 kg^{-1} × 8,360 kg m^{-3} × 3.42 m^3/17.5×10^6 W). Hence, comparison of these raw materials commodity costs alone (that is, neglecting the non-trivial processing costs for forming thin plates of these Ni-based alloys, preparing patterned channels into such plates via photochemical etching, and other heat exchanger manufacturing steps) to the manufacturing (raw materials + processing) cost obtained above for compact ZrC/W-based heat exchangers (<\$0.043 W^{-1}; <\$25.1 kg^{-1}) indicated that the latter heat exchangers would be comparable in price or less expensive than compact heat exchangers comprised of such high-temperature Ni-based alloys, while achieving twice the power density (8.2 $MW_{th}\,m^{-3}$ versus 4.1 $MW_{th}\,m^{-3}$).

The cost of a compact (printed-circuit-type) heat exchanger comprised of 316 SS was also considered. The combined commodity-based cost of the major components of 316 SS (>62 wt% Fe, 17 wt% Cr, 12 wt% Ni, 2.5 wt% Mo) was found to be at least \$3.6 kg^{-1} (with ≤\$0.13 kg^{-1} for Fe, from iron fines at \$0.078 kg^{-1}).

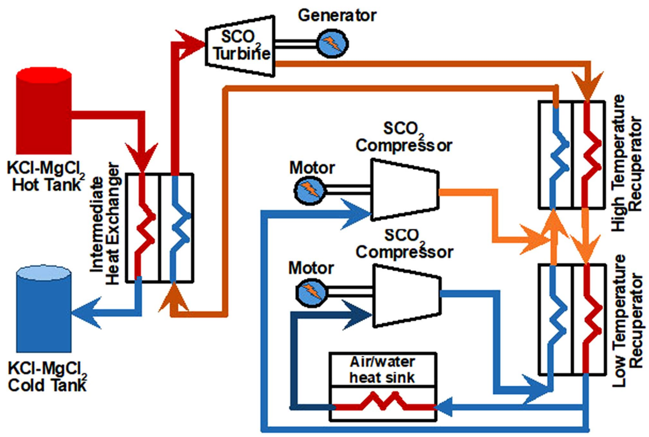
The solid volume of 316 SS in the 17.5 MW_{th} heat exchanger discussed above (with a power density of only up to 1.2 $MW_{th}\,m^{-3}$) with overall dimensions (width × length × height) of 0.60 m × 2.83 m × 8.32 m (comprised of a stack of 2,656 plates, each of 3.12 mm thickness with semi-circular channels of 2 mm diameter and centre-to-centre spacing of 8.30 mm, and 91,200 channels for each fluid) would be 13.3 m^3, that is: $[0.6\ m \times 2.83\ m \times 8.32\ m] - [\pi(0.001\ m)^2 \times 2.83\ m \times 0.5 \times 91{,}200 \times 2]$. For a 316 SS density of 7,990 kg m^{-3}, such a 17.5-MW_{th} IN740H heat exchanger would have mass 1.06×10^5 kg. Hence, the raw materials commodity cost alone of such a 17.5-MW_{th} IN740H heat exchanger would be at least \$0.022 W^{-1} (\$3.6 $kg^{-1} \times 1.06 \times 10^5$ kg/17.5×10^6 W).

The photochemical etching cost of 316 SS heat exchanger plates was then considered. Assuming a conservative photochemical etching cost per area of 316 SS of \$170 m^{-2} (three times lower than the lowest cost of \$520 m^{-2} obtained from vendor quotes for photochemical etching of 2 mm diameter semi-circular channels into 316 SS), the cost of etching 2,656 plates of 2.83 m length and 0.60 m width (total area of plate surfaces to be etched is $4.51 \times 10^3\ m^2$) for a 17.5-MW_{th} heat exchanger was found to be \$0.044 W^{-1} (\$170 $m^{-2} \times 4.51 \times 10^3\ m^2/17.5 \times 10^6$ W). Hence, the combined cost estimate (raw materials commodity cost + etching cost) for the 17.5 MW_{th} 316 SS heat exchanger was found to be at least \$0.066 W^{-1} (that is, ignoring processing costs other than photochemical etching). Such techno-economic analyses indicated that, for a sufficient manufacturing throughput, compact (printed circuit-type) ZrC/W-based heat exchangers can be manufactured at a competitive or lower cost than for state-of-the-art nickel superalloy-based heat exchangers and stainless-steel-based heat exchangers, while achieving at least twice the power density (8.2 $MW_{th}\,m^{-3}$ versus 4.1 $MW_{th}\,m^{-3}$ or 1.2 $MW_{th}\,m^{-3}$).

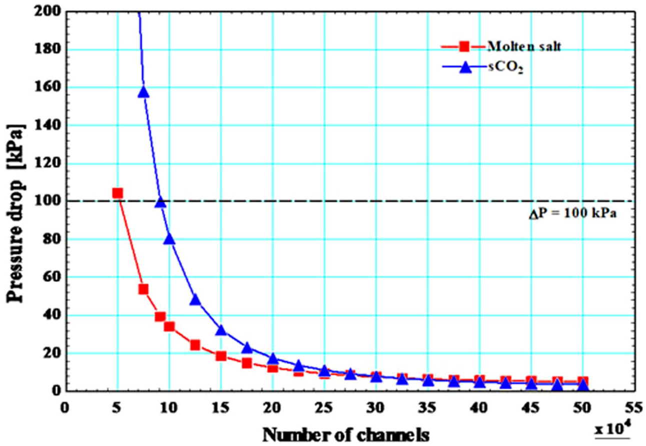
Data availability

Data generated or analysed during this study are available from the corresponding author on reasonable request. Data reported are available within the paper. The Ricardo Excel model of ZrC/W-based heat exchanger processing costs is available from the corresponding author on reasonable request.

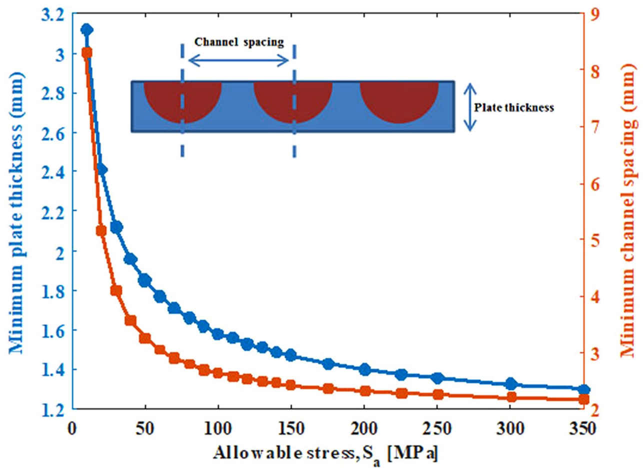
31. Lipke, D. W., Zhang, Y., Liu, Y., Church, B. C. & Sandhage, K. H. Near net shape/net dimension ZrC/W-based composites with complex geometries via rapid prototyping and displacive compensation of porosity (DCP). *J. Eur. Ceram. Soc.* **30**, 2265–2277 (2010).
32. Joint Committee on Powder Diffraction Standards (JCDPS) *JCPDS International Center for Diffraction Data File* 00-004-0836 for Cu, 01-071-3645 for Cu_2O, 00-048-1827 for WO_2 and 01-070-2491 for monoclinic ZrO_2, http://www.icdd.com/index.php/pdf-4/ (JCPDS International Center for Diffraction Data, Newtown Square, 2007).
33. Barin, I. *Thermochemical Data of Pure Substances* 3rd edn (VCH, Weinheim, 1995).
34. Narula, M. L., Tare, V. B. & Worrell, W. L. Diffusivity and solubility of oxygen in solid copper using potentiostatic and potentiometric techniques. *Metall. Trans. B* **14**, 673–677 (1983).
35. Sieverts, A. The absorption of gases by metals. *Z. Metallk.* **21**, 37–46 (1929).
36. Grjotheim, K., Holm, J. L. & Roetnes, M. Phase diagrams of the systems sodium chloride-magnesium chloride and potassium chloride-magnesium chloride. *Acta Chem. Scand.* **26**, 3802–3803 (1972).
37. Sohal, M. S., Ebner, M. A., Sabharwall, P. & Sharpe, P. *Engineering Database of Liquid Salt Thermophysical and Thermochemical Properties.* Technical Report INL/EXT-10-18297 https://www.osti.gov/biblio/980801-engineering-database-liquid-salt-thermophysical-thermochemical-properties (Idaho National Laboratory, Idaho Falls, 2010).
38. Moisseytsev, A., Sienicki, J. J., Cho, D. H. & Thomasa, M. R. Comparison of heat exchanger modeling with data from CO_2-to-CO_2 printed circuit heat exchanger performance tests. In *Proc. Int. Congress Adv. Nucl. Power Plants (ICAPP) 2010* Paper 2284, 459–468, http://www.ans.org/store/item-700358/ (American Nuclear Society, La Grange Park, 2010).
39. American Society of Mechanical Engineers (ASME) *2004 ASME Boiler and Pressure Vessel Code. Section VIII, Division 1* NG-3000 (ASME, New York, 2004).
40. Specialty Metals Corp. *Inconel Alloy 740H A Superalloy Specifically Designed For Advanced Ultra Supercritical Power Generation* http://www.specialmetals.com/assets/smc/documents/alloys/inconel/inconel-alloy-740-h.pdf (Specialty Metals Corp., New Hartford, 2018).
41. Specialty Metals Corp. *Inconel Alloy 617* http://www.specialmetals.com/assets/smc/documents/alloys/inconel/inconel-alloy-617.pdf (Specialty Metals Corp., New Hartford, 2018).



Extended Data Fig. 1 | Schematic illustration of a concentrated solar power plant. The thermal energy storage medium is KCl-$MgCl_2$ molten salt (67% mol%–33 mol%[36,37]) and the plant uses a sCO_2 Brayton cycle for power generation.



Extended Data Fig. 2 | More channels in the heat exchanger reduce the values of pressure drop. Variations of the channel length, and the values of the pressure drop for the molten salt and sCO_2 streams, are plotted as a function of the number of channels for each fluid.



Extended Data Fig. 3 | Higher allowed stresses reduce the required plate thickness and channel spacing. Values of the minimum plate thickness and minimum channel spacing associated with a given allowed stress are plotted for a compact, printed-circuit-type heat exchanger.

Extended Data Table 1 | Characteristics of the reaction for fabrication of the heat exchanger plates

Sample Number	W:WC	$\Delta m/m_o$	Reaction Extent*	Phase Content (Vol %)				Density**	Porosity
	(molar ratio)	(%)	(%)	ZrC	W	WC	Cu	[Exptl:Theo] ($g\cdot cm^{-3}$)	(%)
1	16.4	47.9	94.3	58.1	35.6	2.9	3.4	11.21 : 11.47	2.3
2	15.4	49.9	93.9	56.9	34.9	3.0	5.2	11.14 : 11.43	2.5
3	24.3	49.6	96.0	58.2	35.7	1.9	4.2	11.16 : 11.41	2.2
4	28.3	49.2	96.6	58.7	36.0	1.7	3.6	11.15 : 11.41	2.3
5	29.0	49.3	96.7	58.7	36.0	1.6	3.7	11.11 : 11.41	2.6
Average	**22.7**	**49.2**	**95.5**	**58.1**	**35.7**	**2.2**	**4.0**	**11.15 : 11.43**	**2.4**
S. Deviation	**6.4**	**0.8**	**1.3**	**0.7**	**0.4**	**0.6**	**0.7**	**0.03 : 0.02**	**0.2**

The table shows the values of the molar W:WC ratio (from quantitative X-ray diffraction analysis), mass change upon infiltration, extent of reaction, phase content, density (measured and theoretical) and bulk porosity of five infiltrated and reacted specimens. The standard deviation of the average values are given.
*Reaction extent (%) = 100[(W:WC)/(1 + (W:WC)].
**Experimental density values and theoretical (zero porosity) density values (calculated from the indicated phase content for each converted ZrC/W-based specimen) are provided.

Extended Data Table 2 | Material characteristics for ZrC/W, Inconel 740H and Inconel 617

Material	$T_{solidus}$ (K)	C_p ($J \cdot g^{-1} \cdot K^{-1}$)	κ ($W \cdot m^{-1} \cdot K^{-1}$)	σ_F (MPa)
ZrC/W	3073[17]	0.285[#]	66.0[#]	369[#*]
IN740H	1561[40]	0.573[40]	22.1[40]	56[40**]
IN617	1605[41]	0.611[41]	25.5[41]	54[41***]

The table shows solidus temperatures ($T_{solidus}$), and average values of the high-temperature (1,073 K) specific heat capacity C_p, thermal conductivity κ and failure strength σ_F. Data are from refs [17,40,41], as indicated.

#From the present work.

*Fracture strength.

**Creep rupture stress (100,000 h).

***Creep rupture stress (100,000 h) obtained by interpolation of data at 760 °C and 870 °C.

Extended Data Table 3 | Operating conditions representative of an intermediate heat exchanger for a 10-MW$_e$ concentrated solar power plant

	Molten salt		**SCO$_2$**	
	Pressure *(MPa)*	***Temperature*** *(K)*	***Pressure*** *(MPa)*	***Temperature*** *(K)*
Inlet	0.1013	1073	20	873
Outlet	-	883	-	1063
Mass flowrate *(kg/s)*	80		72.5	

The intermediate (primary) heat exchanger is located between the high-temperature fluid heated by sunlight and the supercritical carbon dioxide in the power block, as shown in Extended Data Fig. 1. $Q = 17.5$ MW$_{th}$ for this molten-salt-to-sCO$_2$ compact heat exchanger.

LETTER

https://doi.org/10.1038/s41586-018-0589-x

Glacial expansion of oxygen-depleted seawater in the eastern tropical Pacific

Babette A. A. Hoogakker[1,2]*, Zunli Lu[3,4]*, Natalie Umling[5], Luke Jones[2], Xiaoli Zhou[6], Rosalind E. M. Rickaby[2], Robert Thunell[5,10], Olivier Cartapanis[7] & Eric Galbraith[8,9]

Increased storage of carbon in the oceans has been proposed as a mechanism to explain lower concentrations of atmospheric carbon dioxide during ice ages; however, unequivocal signatures of this storage have not been found[1]. In seawater, the dissolved gases oxygen and carbon dioxide are linked via the production and decay of organic material, with reconstructions of low oxygen concentrations in the past indicating an increase in biologically mediated carbon storage. Marine sediment proxy records have suggested that oxygen concentrations in the deep ocean were indeed lower during the last ice age, but that near-surface and intermediate waters of the Pacific Ocean—a large fraction of which are poorly oxygenated at present—were generally better oxygenated during the glacial[1–3]. This vertical opposition could suggest a minimal net basin-integrated change in carbon storage. Here we apply a dual-proxy approach, incorporating qualitative upper-water-column and quantitative bottom-water oxygen reconstructions[4,5], to constrain changes in the vertical extent of low-oxygen waters in the eastern tropical Pacific since the last ice age. Our tandem proxy reconstructions provide evidence of a downward expansion of oxygen depletion in the eastern Pacific during the last glacial, with no indication of greater oxygenation in the upper reaches of the water column. We extrapolate our quantitative deep-water oxygen reconstructions to show that the respired carbon reservoir of the glacial Pacific was substantially increased, establishing it as an important component of the coupled mechanism that led to low levels of atmospheric carbon dioxide during the glacial.

The modern-day Pacific Ocean contains a vast volume of oxygen-depleted waters. In the eastern basin north of 18° S, waters deeper than 1 km (deepening to 2 km north of the Equator) are generally oxic (with an oxygen concentration, [O_2], of more than 120 μmol kg^{-1}), whereas at shallower depths most waters are hypoxic ([O_2] < 60–120 μmol kg^{-1}), and a small fraction are suboxic[6] ([O_2] < 2–10 μmol kg^{-1}). The eastern tropical North Pacific (ETNP) oxygen minimum zone (OMZ) is the world's largest OMZ, and currently encompasses 67% of the suboxic waters on Earth[6]. Low-oxygen conditions place important limitations on marine life, with hypoxic conditions proving lethal for more than half of marine benthic animal species[7]. Oceanic nutrient cycling is also affected by suboxic conditions[8,9], under which the remineralization of organic material occurs via anaerobic metabolic pathways, including denitrification and anammox. This removes bioavailable nitrogen (which supports primary production) from the ocean and generates the greenhouse gas nitrous oxide.

Because of the intrinsic link between oxygen and carbon in photosynthesis and respiration, oxygen utilization provides a direct reflection of the strength of the biological carbon pump and therefore its influence on atmospheric CO_2[4]. Today, the Pacific Ocean represents the largest modern sink of respired organic carbon (>730 Gt, around 50% of the global ocean inventory[10]), half of which resides in the upper 1.5 km.

The concentration of dissolved oxygen in seawater is controlled by two factors: first, the saturation oxygen concentration of seawater in contact with the atmosphere, which is the sum of oxygen solubility (a function of temperature and salinity) and any disequilibrium from saturation at the ocean surface; and second, the net oxygen utilization, which is determined by the accumulated consumption during remineralization of organic material along the pathways of advection and mixing[8]. Over the past 50 years the observed vertical expansion of the equatorial Pacific OMZ has been attributed mostly to a net increase in oxygen utilization, which could reflect a reduced input rate of oxygen through advection and mixing and/or an increase in the local rate of respiration by organic matter[11,12]. A further decline in ocean oxygen levels is predicted by Earth system models under anthropogenic warming, linked to increased temperatures (lowering the saturation oxygen concentration) and increased oxygen utilization owing to decreased ventilation[8,11,13]. However, model simulations disagree about oxygen changes in the tropical thermoclines, and do not reproduce the large historical changes[11], which suggests that these models are missing important processes that may compromise their predictions of future change[13,14].

Reconstructions of the last ice age offer an alternative test of the link between climate and ocean oxygenation. Lower glacial seawater temperatures would have increased oxygen saturation concentrations[2] and decreased remineralization rates[15]. These conditions could have resulted in a better-oxygenated upper ocean, potentially eliminating the OMZs. Bulk sedimentary nitrogen isotope ($\delta^{15}N$) records from the eastern tropical Pacific (ETP)[16,17] have been interpreted to reflect overall reduced glacial denitrification rates in the upper water column[18], which could indicate an absence of suboxic waters. By contrast, the cold-enhanced solubility appears to have been overwhelmed by increased oxygen utilization in the deep Pacific, resulting in reduced oxygen concentrations and increased respired carbon storage that could have contributed to the low atmospheric CO_2 concentrations[1–3]. However, these reconstructions are based on qualitative proxies, which are often difficult to interpret[19]. Furthermore, many of these records have been limited to core sites from continental slopes, and are potentially biased by local conditions[19].

To constrain upper-water-column oxygen concentrations, we used planktonic foraminifera I/Ca ratios[5] (see Methods). This proxy takes advantage of iodine speciation in seawater. The iodate species (IO_3^-) is favoured under well-oxygenated settings, whereas iodide (I^-) becomes the dominant species under oxygen-depleted conditions. Because foraminiferal calcite incorporates only iodate, the foraminiferal I/Ca ratio therefore reflects the abundance of the oxidised form[20].

Furthermore, we use the benthic foraminiferal carbon-isotope gradient proxy ($\Delta\delta^{13}C$) to quantitatively reconstruct bottom-water oxygen concentrations[4]. The $\Delta\delta^{13}C$ between bottom water and pore water at the anoxic boundary in sediments is related to the oxygen concentration of the overlying bottom waters[21]. The $\Delta\delta^{13}C$ between

[1]The Lyell Centre, Heriot-Watt University, Edinburgh, UK. [2]Department of Earth Sciences, University of Oxford, Oxford, UK. [3]Department of Earth Sciences, Syracuse University, Syracuse, NY, USA. [4]State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen, China. [5]School of Earth, Ocean and Environment, University of South Carolina, Columbia, SC, USA. [6]Department of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ, USA. [7]University of Bern, Oeschger Centre for Climate Change Research, Bern, Switzerland. [8]Institut de Ciencia i Tecnologia Ambientals (ICTA) and Department of Mathematics, Universitat Autonoma de Barcelona, Bellaterra, Spain. [9]ICREA, Barcelona, Spain. [10]Deceased: Robert Thunell. *e-mail: b.hoogakker@hw.ac.uk; zunlilu@syr.edu

bottom water and pore water at the anoxic boundary is reproduced by the $\Delta\delta^{13}C$ of benthic foraminifera with microhabitats in bottom water (*Cibicidoides wuellerstorfi*) and in sediments at the anoxic boundary (*Globobulimina* spp.)[4]. This method enables us to quantitatively reconstruct past dissolved oxygen concentrations in the range of 55–235 μmol kg^{-1} (see Methods) in bottom waters from tropical to temperate regions, with an estimated total standard error[4] of 17 μmol kg^{-1}. Our tandem proxy approach enables us to place firm constraints on past changes in the geometry of oxygen-depleted waters in the eastern tropical Pacific over the past 40,000 years. Furthermore, extrapolation of our new quantitative bottom-water oxygen reconstructions enables us to calculate the change in size of the Pacific respired-carbon pool and assess its role in glacial–interglacial CO_2 cycles.

Planktonic foraminifera I/Ca ratios were measured at two eastern tropical Pacific sites. ODP site 1242 (7.86° N, 83.61° W, 1.36 km) is on the Costa Rica margin, in the eastern tropical North Pacific (ETNP), whereas ODP site 849 (0.18° N, 110.50° W, 3.85 km) lies beneath the eastern equatorial cold tongue (Fig. 1). Planktonic foraminifera I/Ca ratios at the ETNP site are expected to monitor changes in the upper boundary of the ETNP-OMZ. The cold tongue site, ODP site 849, is distal from modern suboxic zones but downstream of waters that have passed through them, and planktonic foraminifera I/Ca ratios at this location are expected to have responded to the broader presence of oxygen-depleted waters within the ETP-OMZ. The location of ODP site 1242 at the deep boundary of the present-day ETNP-OMZ is ideal to test for changes in the vertical extent of the OMZ, via benthic foraminifera $\Delta\delta^{13}C$. Additionally, bottom-water oxygen concentrations were reconstructed for deep water at TR163-25 (1.65° S, 88.45° W, 2.65 km), to provide quantitative estimates of changes in deep-water oxygen concentrations in the eastern tropical Pacific and calculate the glacial increase in the deep Pacific respired-carbon pool. Details of age models are provided in Extended Data Tables 1, 2 and Extended Data Fig. 1.

Modern oxygen profiles at ODP sites 849 and 1242 are very similar (Fig. 1), except that OMZ waters ([O_2] threshold[22] < 45 μmol kg^{-1}) occur at a much shallower depth at the ETNP site (within the upper 50 m) compared to the cold tongue site (deeper than 250 m) (Fig. 1). This difference in the upper water column is consistent with the contrasting core–top planktonic foraminifera I/Ca values at the two sites (Fig. 2). If suboxia had been reduced during the glacial, as has been previously suggested, one would expect high I/Ca values to be found in glacial-age foraminifera. Instead we find that low I/Ca values (<0.6 μmol mol^{-1}) prevailed continuously over the past 40 thousand years (kyr) at the ETNP-OMZ site, which is consistent with persistent oxygen depletion at shallow depths (Fig. 2). Furthermore, although I/Ca ratios of all planktonic species in the cold tongue from 40–25 kyr before present (BP) were similar to values from the late Holocene, during early deglaciation (around 18–16 kyr BP) the I/Ca of shallow-dwelling species decreased to values as low as those of the thermocline species. The persistently depleted planktonic foraminifera oxygen isotope values of the shallow-dwelling species and the heavy values of the thermocline species (Fig. 2) indicate similar depth habitats over the past 40 kyr. Therefore, we attribute the lower I/Ca values of the shallow-dwelling species at site 849 during early deglaciation to the increased presence of oxygen-depleted waters in the ETP-OMZ.

Turning to the deep sea, reconstructed dissolved oxygen at ODP site 1242 shows generally lower concentrations during the glacial compared to the Holocene, with an average Last Glacial Maximum (LGM) (18–22 kyr BP) dissolved oxygen content of 55 μmol kg^{-1} (±17 μmol kg^{-1}, Fig. 3). The lowest oxygen concentrations (44 μmol kg^{-1}) were recorded during early deglaciation (17–15 kyr BP), followed by a rapid increase in the mid- to late deglaciation. Maximum oxygen concentrations of 100 μmol kg^{-1} were recorded during the early Holocene. Oxygenation then decreased slightly through the Holocene, reaching late Holocene values of 85 μmol kg^{-1} (Fig. 3). At the deeper site TR163-25, reconstructed LGM oxygen concentrations are similar to those of ODP site 1242, averaging 54 μmol kg^{-1} (Fig. 3), and there is also a brief decline in dissolved oxygen during the early deglaciation to around 40 μmol kg^{-1}, followed by a rapid increase to around 160 μmol kg^{-1} in the mid-Holocene (Fig. 3).

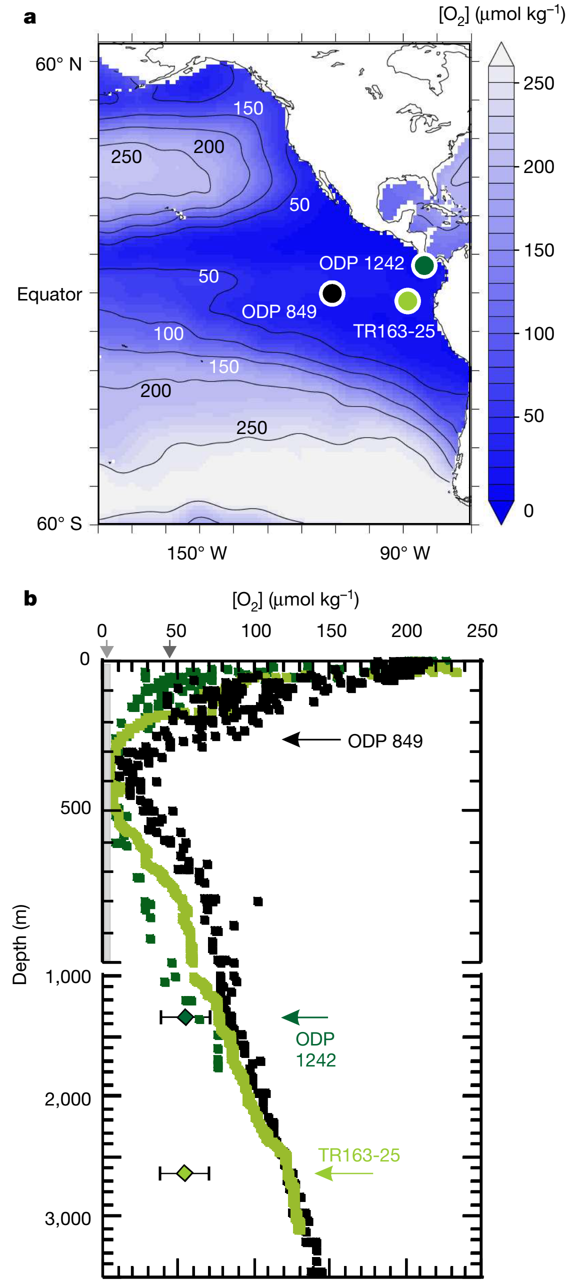


Fig. 1 | Overview of dissolved oxygen concentrations in the eastern Pacific Ocean. a, Oxygen concentrations between 60° S and 60° N at 400 m water depth (circles show core locations). Data are from ref. [28]. **b**, Vertical profiles at the core sites (from https://www.nodc.noaa.gov/OC5/SELECT/dbsearch/dbsearch.html; data from ref. [28]). ODP site 1242, dark green; ODP site 849, black; TR163-25, light green. Note the different scales for the upper part (0–1,000 m) and the lower part (1,000–4,000 m) of the water column. Arrows on the *x* axis indicate [O_2] thresholds for suboxia (light grey) and the OMZ (dark grey). Diamonds illustrate the reconstructed LGM bottom-water [O_2] values at ODP site 1242 and TR163-25, including ±17 μmol kg^{-1} error[4].

Our dual-proxy results from the upper 1.4 km of the water column (planktonic foraminifera I/Ca at ODP sites 1242 and 849, $\Delta\delta^{13}C$ at ODP site 1242) show sustained oxygen depletion; this is in contrast with other studies, which have suggested that the upper water column in the Pacific was generally more oxygenated at this time[1–3]. These previous conclusions were based on observations of low sedimentary $\delta^{15}N$ (interpreted as lower rates of denitrification), weaker sedimentary

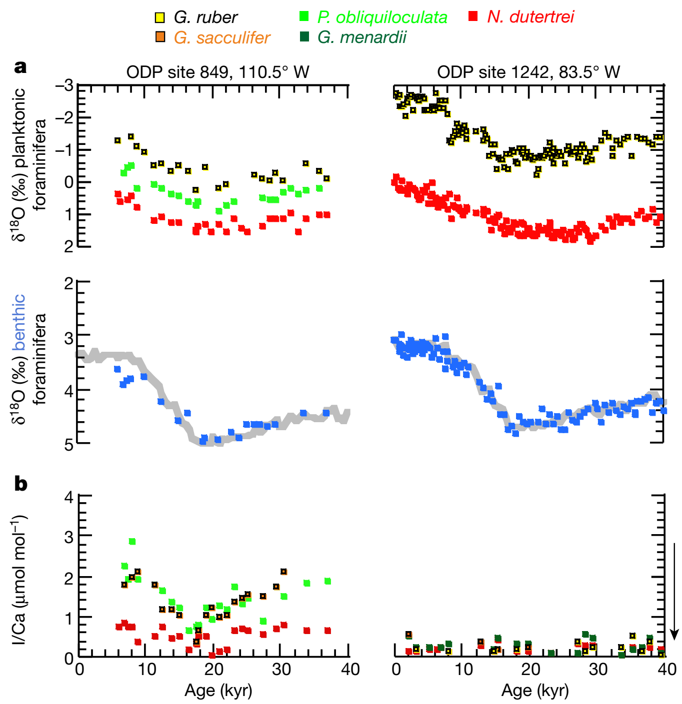


Fig. 2 | Reconstructed ETP surface water oxygenation. a, Planktonic foraminiferal and benthic composite oxygen isotope ($\delta^{18}O$) records (blue symbols) and stacked records (grey lines)[29] at ODP sites 849 and 1242. Planktonic foraminiferal oxygen isotopes at ODP site 1242 until 28 kyr BP are from ref. [30]. Details of age models can be found in Methods. **b**, I/Ca ratios of planktonic foraminifera. I/Ca ratios of less than 2.5 μmol mol^{-1} are indicative of the presence of low-oxygen waters in the upper 400 m of the water column[5]. The arrow indicates the increasing influence of oxygen depletion.

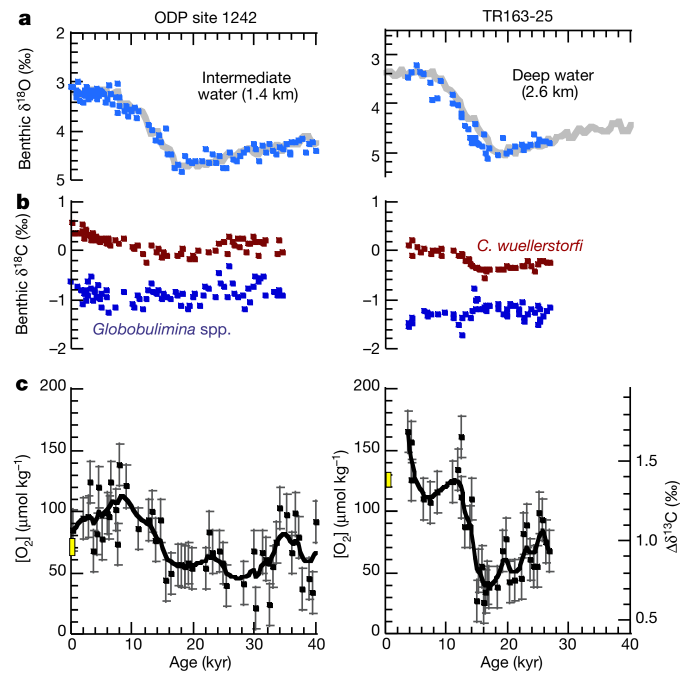


Fig. 3 | Reconstructed ETP bottom-water oxygen concentrations. a, Benthic foraminiferal $\delta^{18}O$ (blue symbols) of ODP site 1242 and TR163-25 (*C. wuellerstorfi*, adjusted by +0.64‰) and stacked records (grey lines) from the intermediate and deep Pacific[29]. Details of age models can be found in Methods. **b**, Benthic foraminiferal carbon isotopes of *C. wuellerstorfi* (red) and *Globobulimina* spp. (blue). **c**, Reconstructed bottom-water $[O_2]$ and $\Delta\delta^{13}C$ (raw data[4], black squares + total error of ±17 μmol kg^{-1}; thick line shows moving average calculated using the boxcar algorithm). Yellow boxes indicate the modern range of bottom-water oxygen concentrations.

laminations and lower abundances of oxygen-sensitive trace metals during the glacial[2]. However, there are several reasons that sedimentary $\delta^{15}N$ could have been lower during the glacial without a substantial change in oxygen concentrations (see Methods and Extended Data Fig. 2). Furthermore, the sedimentary laminations and trace metals previously examined at three sites in the coastal ETP showed only weak signs of oxygen change between the LGM and the Holocene[16,17], which could also be attributed to changes in the characteristics of accumulating sediments[23,24]. Therefore, the persistently low I/Ca values, in combination with reduced glacial bottom-water oxygen levels at 1.4 km (today the lower boundary of the ETNP-OMZ), do not support a substantial contraction of the upper reaches of the tropical Pacific OMZ during the glacial period compared to today.

Our results also indicate a period of particularly strong oxygen depletion during the early deglaciation, which is consistent with previous sedimentary $\delta^{15}N$ values, lamination, and trace metal evidence from the ETNP[16,17]. The convergence of mixed-layer and thermocline planktonic foraminifera to low values of I/Ca at ODP site 849 (Fig. 2) suggests that the downward expansion of oxygen-depleted waters in the ETP-OMZ, indicated by the bottom-water oxygen reconstructions (Fig. 3), was accompanied by an intensified influence of oxygen depleted-waters in the upper water column. The interval coincides with a weak Atlantic Meridional Overturning Circulation, and an apparent productivity peak in the eastern equatorial Pacific that is speculated to reflect an increased delivery of nutrients from southern-sourced deep waters and intensified upwelling[17,25–27].

Our tandem proxy results provide new insights into the evolution of respired carbon storage in the eastern tropical Pacific since the last ice age. Today, a quarter of the total global respired carbon reservoir is stored in the upper 1.5 km (intermediate and subsurface waters) of the Pacific. Our results suggest that the respired-carbon reservoir of the upper water column has shown little change between the LGM and the Holocene, whereas that of the deeper Pacific has increased, suggesting a net increase in the size of the Pacific glacial respired-carbon pool.

Furthermore, the results of $\Delta\delta^{13}C$ analysis show that the modern vertical-oxygen gradient ($\Delta[O_2]$, of around 65 μmol kg^{-1}) between water depths of 1.4 km and 2.6 km was eliminated during the LGM (Fig. 1), so that oxygen concentrations did not increase with depth as they do today. We also find that the gradient of $\delta^{13}C$ in the dissolved inorganic carbon between these water masses was reversed (Extended Data Fig. 3), as would be expected given the respired carbon concentrations inferred from our quantitative oxygen reconstructions and similar changes in the preformed component of $\delta^{13}C$ (for details, see Methods). Our data therefore suggest that, despite large changes in the average $\delta^{13}C$ of dissolved inorganic carbon for the whole ocean and changes in air–sea exchange, the relative change in $\delta^{13}C$ between sites in the depth range of 1.4 km to 3 km provides a good approximation of the change in oxygen concentrations.

We take advantage of this new constraint, together with our LGM–modern $\delta^{13}C$ compilation, to extrapolate our results spatially in the deep Pacific. Our results suggest that the total amount of respired carbon in the Pacific was approximately 90 Gt greater between water depths of 1.4 km and 3 km, and possibly 200 Gt greater across the whole of the deep Pacific (see Methods), during the LGM compared with today. This provides a useful new target for model simulations of glacial carbon cycling. Although the average increase in respired carbon concentrations in deeper waters of the Pacific is only half that of the deep Atlantic[4], the estimated glacial increase in its respired carbon reservoir is almost three times that of the deep Atlantic owing to its vast size. This suggests that the Pacific made an important contribution to glacial–interglacial changes in atmospheric CO_2 levels.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0589-x.

Received: 3 July 2017; Accepted: 13 August 2018;
Published online 17 October 2018.

1. Sigman, D. M. & Boyle, E. A. Glacial/interglacial variations in atmospheric carbon dioxide. *Nature* **407**, 859–869 (2000).
2. Galbraith, E. D. & Jaccard, S. L. Deglacial weakening of the oceanic soft tissue pump: global constraints from sedimentary nitrogen isotopes and oxygenation proxies. *Quat. Sci. Rev.* **109**, 38–48 (2015).
3. Bradtmiller, L. I., Anderson, R. F., Sachs, J. P. & Fleisher, M. Q. A deeper respired carbon pool in the glacial equatorial Pacific Ocean. *Earth Planet. Sci. Lett.* **299**, 417–425 (2010).
4. Hoogakker, B. A. A., Elderfield, H., Schmiedl, G., McCave, I. N. & Rickaby, R. E. M. Glacial–interglacial changes in bottom-water oxygen content on the Portuguese margin. *Nat. Geosci*. **8**, 40–43 (2015).
5. Lu, Z. et al. Oxygen depletion recorded in upper waters of the glacial Southern Ocean. *Nat. Commun*. **7**, 11146 (2016).
6. Bianchi, D., Dunne, J. P., Sarmiento, J. L. & Galbraith, E. D. Data-based estimates of suboxia, denitrification, and N_2O production in the ocean and their sensitivities to dissolved O_2. *Global Biogeochem. Cycles* **26**, (2012).
7. Vaquer-Sunyer, R. & Duarte, C. M. Thresholds of hypoxia for marine biodiversity. *Proc. Natl Acad. Sci. USA* **105**, 15452–15457 (2008).
8. Keeling, R. E., Körtzinger, A. & Gruber, N. Ocean deoxygenation in a warming world. *Ann. Rev. Mar. Sci*. **2**, 199–229 (2010).
9. Lam, P. & Kuypers, M. M. M. Microbial nitrogen cycling processes in oxygen minimum zones. *Ann. Rev. Mar. Sci*. **3**, 317–345 (2011).
10. Schmittner, A. & Somes, C. J. Complementary constraints from carbon (^{13}C) and nitrogen (^{15}N) isotopes on the glacial ocean's soft-tissue biological pump. *Paleoceanography* **31**, 669–693 (2016).
11. Schmidtko, S., Stramma, L. & Visbeck, M. Decline in global oceanic oxygen content during the past five decades. *Nature* **542**, 335–339 (2017).
12. Stramma, L., Johnson, G. C., Sprintall, J. & Mohrholz, V. Expanding oxygen-minimum zones in the tropical oceans. *Science* **320**, 655–658 (2008).
13. Bopp, L. et al. Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models. *Biogeosciences* **10**, 6225–6245 (2013).
14. Long, M., Deutsch, C. & Ito, I. Finding forced trends in oceanic oxygen. *Global Biogeochem. Cycles* **30**, 381–397 (2016).
15. Matsumoto, K. Biology-mediated temperature control on atmospheric pCO_2 and ocean biogeochemistry. *Geophys. Res. Lett.* **34**, L20605 (2007).
16. Pichevin, L. E. et al. Interhemispheric leakage of isotopically heavy nitrate in the eastern tropical Pacific during the last glacial period. *Paleoceanography* **25**, PA1204 (2010).
17. Hendy, I. L. & Pedersen, T. F. Oxygen minimum zone expansion in the eastern tropical North Pacific during deglaciation. *Geophys. Res. Lett.* **33**, L20602 (2006).
18. Galbraith, E. D., Kienast, M. & The NICOPP working group members. The acceleration of ocean denitrification during deglacial warming. *Nat. Geosci*. **6**, 579–584 (2013).
19. Moffitt, S. E. et al. Paleoceanographic insights on recent oxygen minimum zone expansion: lessons for modern oceanography. *PLoS ONE* **10**, e0115246 (2015).
20. Lu, Z., Jenkyns, H. C. & Rickaby, R. E. M. Iodine to calcium ratios in marine carbonates as a paleo-redox proxy during oceanic anoxic events. *Geology* **38**, 1107–1110 (2010).
21. McCorkle, D. C. & Emerson, S. R. The relationship between pore water carbon isotopic composition and bottom water oxygen concentration. *Geochim. Cosmochim. Acta* **52**, 1169–1178 (1988).
22. Karstensen, J., Stramma, L. & Visbeck, M. Oxygen minimum zones in the eastern tropical Atlantic and Pacific oceans. *Prog. Oceanogr*. **77**, 331–350 (2008).
23. van Geen, A. et al. On the preservation of laminated sediments along the western margin of North America. *Paleoceanography* **18**, 1098 (2003).
24. Nameroff, T. J., Calvert, E. & Murray, J. W. Glacial–interglacial variability in the eastern tropical North Pacific oxygen minimum zone recorded by redox-sensitive trace metals. *Paleoceanography* **19**, PA1010 (2004).
25. Costa, K. M. et al. Productivity patterns in the Equatorial Pacific over the last 30,000 years. *Global Biogeochem. Cycles* **31**, 850–865 (2017).
26. Kienast, M. et al. Eastern Pacific cooling and Atlantic overturning circulation during the last deglaciation. *Nature* **443**, 846–849 (2006).
27. de la Fuente, M., Skinner, L., Calvo, E., Pelejero, C. & Cacho, I. Increased reservoir ages and poorly ventilated deep waters inferred in the glacial Eastern Equatorial Pacific. *Nat. Commun*. **6**, 7420 (2015).
28. Garcia, H. et al. *World Ocean Atlas 2013, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation* (ed. Levitus, S.) (NOAA Atlas NESDIS 75, 2013).
29. Stern, J. V. & Lisiecki, L. E. Termination 1 timing in radiocarbon-dated regional benthic $\delta^{18}O$ stacks. *Paleoceanography* **29**, 1127–1142 (2014).
30. Benway, H. M., Mix, A. C., Haley, B. A. & Klinkhammer, G. P. Eastern Pacific warm pool paleosalinity and climate variability: 0–30 kyr. *Paleoceanography* **21**, PA3008 (2006).

Acknowledgements This study benefited from discussions with R. Ganeshram. This work is supported by UK Natural Environment Research Council (NERC) grant NE/I020563/1 (to B.A.A.H.), National Science Foundation (NSF) grants OCE-1232620 and OCE-1736542 (to Z.L.) and Swiss National fund PP00P2_144811 (to O.C.). This research used samples and/or data provided by the Ocean Drilling Program (ODP). ODP is sponsored by the US National Science Foundation and participating countries (Natural Environment Research Council in the UK) under the management of Joint Oceanographic Institutions (JOI), Inc. M. Hall, J. Rolfe and C. Day are acknowledged for help with stable isotope analyses.

Author contributions B.A.A.H. and Z.L. conceived and coordinated the work. B.A.A.H., Z.L., N.U., L.J. and X.Z. carried out data analyses; O.C. carried out data synthesis. B.A.A.H., Z.L. and E.G. constructed the figures and wrote the paper, with contributions from the other co-authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0589-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to B.A.A.H. or Z.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Analytical methods. Foraminifera oxygen and carbon isotopes for ODP sites 849 and 1242 were measured using a Thermo MAT253 IRMS coupled to a Kiel Device at the Godwin Laboratory (University of Cambridge) and a Thermo Delta V Advantage coupled to a Kiel Device at the Department of Earth Sciences (University of Oxford). Calibration to Vienna Pee Dee Belemnite was via NBS19 standards. Overall precision for $\delta^{18}O$ is $\sigma = 0.07‰$ (Oxford) and $\sigma = 0.08‰$ (Cambridge), and for $\delta^{13}C$ is $\sigma = 0.04‰$ (Oxford) and $\sigma = 0.06‰$ (Cambridge). For benthic foraminifera analyses we typically used 3–5 specimens of *C. wuellerstorfi*, 6 specimens of *C. pachyderma*, and >4 specimens of *Globobulimina* spp. For planktonic foraminifera analyses a minimum of 20 specimens were analysed. For site TR163-25 benthic foraminifera, oxygen and carbon isotopes, as well as (homogenized) bulk sedimentary nitrogen isotopes, were measured on a GV Isoprime stable isotope ratio mass spectrometer at the University of South Carolina, with a long-term laboratory reproducibility of 0.07‰ (oxygen) 0.06‰ (carbon), and 0.14‰ (nitrogen). Typically 1–5 *Globobulimina* spp. and *C. wuellerstorfi* were used for benthic foraminifera stable isotope analyses at site TR163-25.

Planktonic foraminifera I/Ca ratios were measured by quadrupole ICP-MS (Bruker M90) at Syracuse University, using a previously published method[5]. The sensitivity of iodine was tuned to above 80 kcps for a 1 p.p.b. standard. Iodine calibration standards were freshly prepared from KIO_3 powder. The precision for ^{127}I is typically better than 1%. The detection limit of I/Ca is on the order of 0.1 μmol mol^{-1}.

Age models. The age models for ODP sites 849 and 1242 are based on oxygen-isotope stratigraphy, matching new benthic foraminiferal $\delta^{18}O$ records (Extended Data Fig. 1, Extended Data Table 1) to the Pacific intermediate and deep-stacked $\delta^{18}O$ records of ref. [29]. The benthic composite $\delta^{18}O$ record of ODP site 849 features specimens of *C. wuellerstorfi*, *Laticarinina pauperata* (both adjusted by +0.64‰ to bring them closer to values of *Uvigerina* spp.), and *Uvigerina* spp. The composite record of ODP site 1242 $\delta^{18}O$ includes mainly specimens of *C. wuellerstorfi*, *Cibicidoides pachyderma* (both adjusted by +0.64‰), and minor contributions from *Uvigerina peregrina*.

For TR163-25 the chronology was developed using one *G. ruber* and three *N. dutertrei* ^{14}C ages (Extended Data Table 2) calibrated with reservoir ages calculated for the EEP from TR163-23[31] and ODP site 1240[27] using the Bayesian age model program BACON[32].

Bottom-water oxygen concentrations. It has been shown[4] that there is a strong ($R^2 = 0.94$) linear relationship between bottom-water oxygen concentrations and $\Delta\delta^{13}C$ at oxygen levels between 55 and 235 μmol kg^{-1}, with an approximately 0.4‰ increase in $\Delta\delta^{13}C$ for every 50 μmol kg^{-1} increase in bottom-water oxygen concentrations. According to ref. [4], the total error associated with bottom-water oxygen concentration at mid- to low latitudes is ± 17 μmol kg^{-1}. When oxygen concentrations exceed 255 μmol kg^{-1}, the relationship with $\Delta\delta^{13}C$ weakens owing to $\delta^{13}C$ of *Globobulimina* spp. becoming much more depleted. This typically occurs in environments in which the oxygen penetration depth is greater than the depth of the sediment mixed layer causing the addition of light carbon through sulfate reduction[21]. At oxygen concentrations between 50 and 20 μmol kg^{-1} we expect the strong linear relationship ($\Delta\delta^{13}C = 0.00772 \times$ (dissolved oxygen concentration) + 0.41446) to hold, as aerobic respiration still dominates the remineralization of organic carbon[33]. This is supported by two new data points derived from temperate North Pacific Holocene samples of ODP sites 1014 ($[O_2] = 32 \pm 10$ μmol kg^{-1}; $\Delta\delta^{13}C = 0.54‰ \pm 0.03‰$) and 1019 ($[O_2] = 21 \pm 6$ μmol kg^{-1}; $\Delta\delta^{13}C = 0.44‰ \pm 0.1‰$). At ODP site 1242, one data point from around 38 kyr BP fell outside of the calibration (reconstructed $[O_2]$ of 16 μmol kg^{-1}) and is not shown in Fig. 3. At ODP site 1242, products of manganese and iron reduction (Mn^{2+} and Fe^{2+}) become important below 50 m composite depth[34] (reconstructions of $\Delta\delta^{13}C$ only took place between 0 and 6.5 m). Therefore, we do not expect deviations in $\Delta\delta^{13}C$ in relation to these processes. The most recent Holocene is missing from core 1242, as evidenced by high core top $\delta^{13}C$ of *C. wuellerstorfi* (average 0.4‰ top 25 cm) in contrast with seawater $\delta^{13}C$ of dissolved inorganic carbon (DIC) of −0.2‰ to −0.3‰[35]. At TR163-25 the late Holocene (<3,500 years) is missing.

Subsurface water oxygen concentrations. To document upper-ocean oxygenation, we use the planktonic foraminifera I/Ca proxy from ref. [5]. The electrode potential of the iodate/iodide couple is very similar to that of denitrification[9]. In the surface ocean, iodide exists in well-oxygenated settings, which has been attributed to disequilibrium caused by biological activity and photochemical reduction of iodate to iodide[36–38]. The oxidation of iodide back to iodate is slow and may take from months to up to 40 years[20].

I/Ca ratios were measured on several planktonic foraminifera species covering a range of depth habits. Spinose species *Globigerinoides sacculifer* (ODP sites 849 and 1242) and *G. ruber* (ODP site 1242) typically live in the surface mixed layer, whereas non-spinose species *Pulleniatina obliquiloculata* (ODP site 849), *Globorotalia menardii* (ODP site 1242) and *Neogloboquadrina dutertrei* (ODP sites 849 and 1242) live deeper, at or below the thermocline[39–41]. These depth habitat differences are expressed in the oxygen isotope records, with consistently depleted values for the warmer surface-mixed-layer species, and heavier values for the deeper- and cooler-water-dwelling species (Fig. 2). Pristine planktonic foraminifera were rigorously cleaned using a previously published method[42] before I/Ca analyses.

It is unlikely that lower deglacial I/Ca ratios at ODP site 849 are due to productivity changes; modern open ocean productivity pulses do not lower IO_3^- to concentrations below 0.25 μM in oxygenated water, suggesting that our planktonic foraminifera I/Ca signals are most likely driven by the oxygen concentration of subsurface water and not by productivity[5].

Nitrogen isotopes. Bulk sedimentary $\delta^{15}N$ can indirectly reflect the extent of suboxia within the upper water column, near the core site, owing to the enrichment of ^{15}N in residual nitrate during denitrification[43]. Nitrogen isotopes can, however, also be affected by other processes such as dilution of the isotopic signal given the fraction of nitrate consumed by denitrification in suboxic zones[44], the input of nitrate by advection from distant suboxic zones[16], the addition of low ^{15}N nitrogen by N_2 fixation, and partial nitrate uptake by phytoplankton at remote locations[18,45], and so are not unambiguous recorders of the local extent of suboxia.

Bulk sedimentary $\delta^{15}N$ at both ODP site 1242 and TR163-25 (Extended Data Fig. 2) show lower values during the LGM, consistent with other $\delta^{15}N$ records within the region[18]. Only at ODP site 1242 are sufficiently low oxygen concentrations ($[O_2] < 2$–4 μmol kg^{-1}) found for denitrification to occur today[46], and only at depths of more than 300 m in the water column (Fig. 1). This is below the depth from which wind-driven upwelling draws. Thus, the nitrogen incorporated in organic matter at the surface and exported to depth, producing the bulk sedimentary $\delta^{15}N$ record, does not directly reflect local suboxia at either site. Instead, the records at these locations are likely to reflect regional changes in nitrogen cycling, as is true for the similar records found throughout the ETP[18]. These changes could have included lower rates of denitrification despite similar volumes of OMZ waters, or more complete nitrate consumption during denitrification leading to a weaker isotopic signal.

Notably, nitrogen isotope values at the Gulf of Tehuantepec, where the most active water column denitrification occurs today, were similar during the LGM and the late Holocene (7‰), consistent with similarly active denitrification during both times[17].

Changes in the soft tissue pump. The $\delta^{13}C$ value of dissolved inorganic carbon ($\delta^{13}C_{DIC}$) depends on both the preformed component ($\delta^{13}C_{pre}$) and soft tissue components ($\delta^{13}C_{soft}$). The latter term results from the remineralization of organic matter and is related through stoichiometric ratios to oxygen consumption and carbon storage. The $\delta^{13}C_{pre}$ is determined by temperature, salinity, p_{CO_2}, alkalinity, the whole ocean average $\delta^{13}C$, and the disequilibrium of surface waters when they sink. Often overlooked, the $\delta^{13}C_{pre}$ value is sensitive to changes in the soft tissue pump and ocean circulation in addition to globally averaged $^{13}C/^{12}C$.

If we ignore the small impact of the carbonate pump on carbon isotopes, the $\delta^{13}C_{DIC}$ at an arbitrary point in the ocean interior is given by:

$$\delta^{13}C_{DIC} = \frac{\delta^{13}C_{pre} \times DIC_{pre} + \delta^{13}C_{soft} \times DIC_{soft}}{DIC_{tot}}$$

The LGM–Holocene change (D) in all quantities is approximately:

$$D\delta^{13}C_{DIC(LGM-Hol)} = \frac{D(\delta^{13}C_{pre} \times DIC_{pre})}{DIC_{tot}} + \frac{D(\delta^{13}C_{soft} \times DIC_{soft})}{DIC_{tot}}$$

This equation includes a number of unknowns, which can be simplified using three assumptions. First, that changes in $\delta^{13}C_{soft}$ were negligible. Second, that although the shallow and deep sites certainly would have had different preformed components, the glacial–interglacial change in the preformed component, $D(\delta^{13}C_{pre} \times DIC_{pre})$, was the same at the two sites. Third, that the change in DIC_{soft}/DIC_{tot} was small. This then gives the change in $\delta^{13}C_{pre}$ between the two depths in (z2–z1) as:

$$D\delta^{13}C_{DIC(z2-z1)} = \delta^{13}C_{soft} \times \frac{DDIC_{soft(z2-z1)}}{DIC_{tot}}$$

The $\delta^{13}C_{DIC}$ data show the relative change between the deep and shallow site, from 0.2‰ during recent times to −0.3‰ during the LGM, a change of 0.5‰. Assuming $\delta^{13}C_{soft}$ is −23‰ and DIC is about 2,200,

$$-0.5 = -23 \times \frac{DDIC_{soft}}{2{,}200} \quad \text{and} \quad DDIC_{soft} = 48$$

This would suggest a glacial–interglacial relative change in oxygen utilization between the two depths of $48 \times 140[O_2]/106\,C = 63$ μM. Our new reconstructions show that oxygen concentrations at the two depths converged at the LGM.

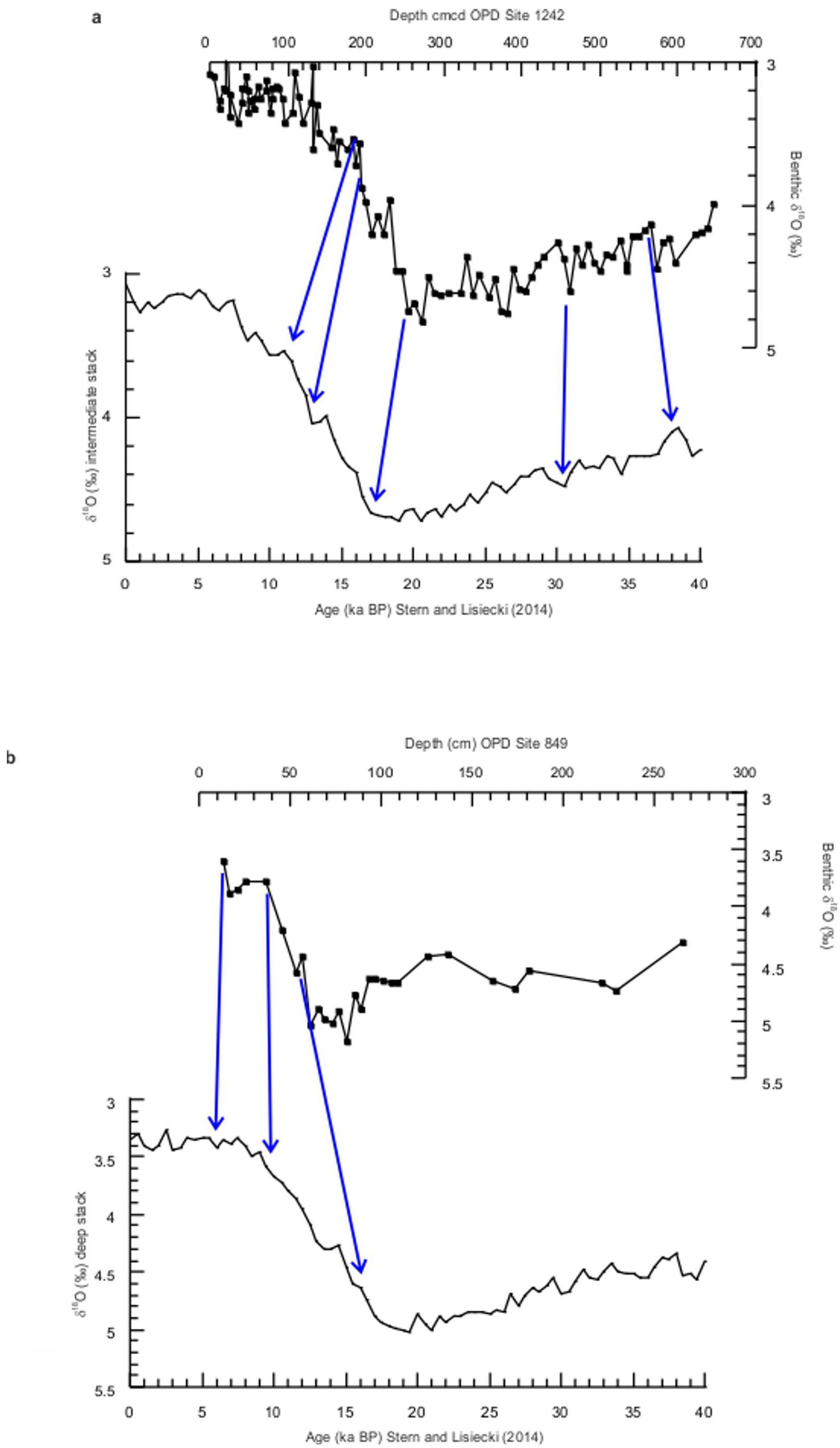
At present, oxygen concentrations at the deeper site are about 65 μM higher than at the shallow site, which would suggest that, on the basis of the $\delta^{13}C$, oxygen concentrations during the LGM should have been the same at the two sites. This is essentially what we observe, supporting the assumption of similar changes in the preformed components in the waters bathing the two depths. Note that this is not to say that the preformed components were constant. Rather, they both changed considerably, but in a coordinated way, owing to the whole ocean change of 0.34‰, and complex interconnected changes in temperature, alkalinity, salinity, p_{CO_2} and air–sea exchange dynamics. Because those changes appear to have occurred together at these depths, we can then take the assumption that, for the Pacific at depths between approximately 1 km and 3 km, there was a uniform LGM–recent change in $\delta^{13}C_{pre}$. As a result, the relative changes in $\delta^{13}C$ between sites should have been dominated by changes in DIC_{soft}, enabling a large-scale budget to be constructed.

Between depths of 1.4 and 3 km, the average difference in $\delta^{13}C$ of DIC between the LGM and recent times is −0.10‰ ± 0.13‰. At TR163-25, LGM–recent $\delta^{13}C$ was −0.30‰, whereas dissolved oxygen values were decreased by 65 μmol kg^{-1} compared with recent times (Extended Data Fig. 3). Thus, with our new constraints, the average decrease of 0.10‰ in the LGM–recent $\delta^{13}C$ of DIC between 1.4 and 3 km in the Pacific can be translated to oxygen concentrations that were 22 μmol kg^{-1} lower (−0.10/0.30 × 65) than preindustrial (not accounting for changes in preformed oxygen disequilibrium). Assuming a 2.5 °C decrease in average deep Pacific temperature and a 1 unit increase in salinity (see ref. [47]), the saturated dissolved oxygen concentration (calculated using the equations in ref. [48]) would be 353 μmol kg^{-1}, nearly 20 μmol kg^{-1} higher than at present. Apparent oxygen utilization (difference between saturation oxygen concentration and measured oxygen concentration) was therefore increased by 42 μmol kg^{-1} during the LGM in the deep Pacific. Extrapolated across water depths between 1.4 and 3 km, this amounts to an increase in respired carbon of 90 Gt C. If similar conditions and changes in $\delta^{13}C_{pre}$ applied across the whole of the deep Pacific (all depths >1.4 km), a volume over which the average LGM–recent $\delta^{13}C$ is −0.17‰ ± 0.18‰, then the corresponding increase in glacial respired carbon would amount to 200 Pg C.

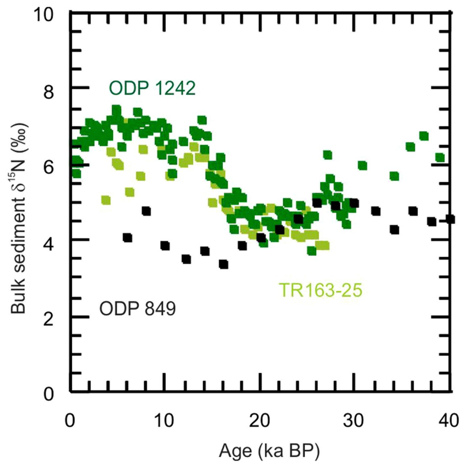
Data availability

Data generated during this study are available from https://doi.pangaea.de/10.1594/PANGAEA.891185.

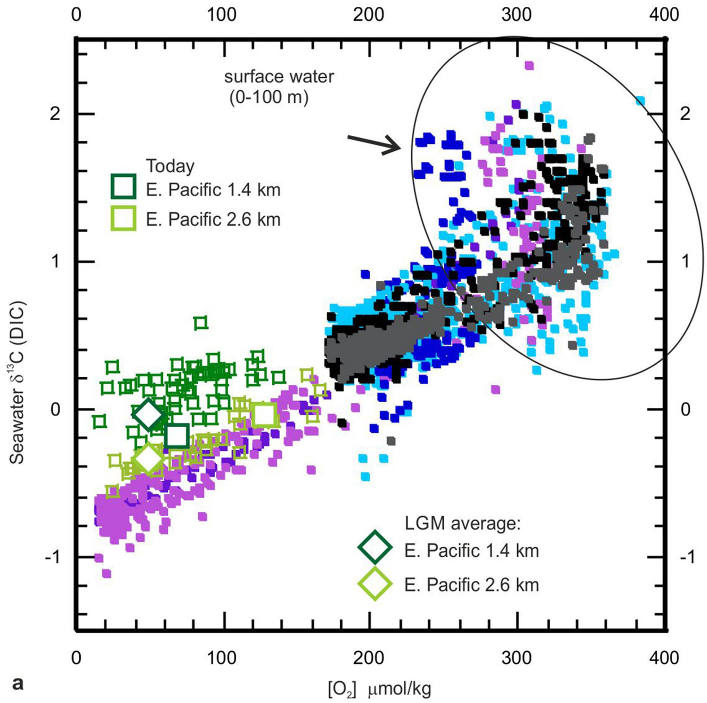
31. Umling, N. E. & Thunell, R. C. Synchronous deglacial thermocline and deep-water ventilation in the eastern equatorial Pacific. *Nat. Commun.* **8**, 14203 (2017).
32. Blaauw, M. & Christen, J. A. Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Anal.* **6**, 457–474 (2011).
33. Codispotti, L., Yoshinari, T. & Devol, A.H. in *Respiration in Aquatic Ecosystems* (eds del Giorgio, P. & Williams, P.) Ch. 12 (Oxford Univ. Press, Oxford, 2005).
34. Mix, A. C. et al. *Proc. ODP, Init. Rep.* https://doi.org/10.2973/odp.proc.ir.202.113.2003 (2003).
35. Eide, M., Olsen, A., Ninnemann, U. S. & Eldevik, T. A global estimate of the full oceanic ^{13}C Suess effect since the preindustrial. *Global Biogeochem. Cycles* **31**, 492–514 (2017).
36. Chance, R. et al. Seasonal and interannual variation of dissolved iodine speciation at a coastal Antarctic site. *Mar. Chem.* **118**, 171–181 (2010).
37. Spokes, L. J. & Liss, P. L. Photochemically induced redox reactions in seawater, II. Nitrogen and iodide. *Mar. Chem.* **54**, 1–10 (1996).
38. Chance, R., Baker, A. R., Carpenter, L. & Jickells, T. D. The distribution of iodide at the sea surface. *Environ. Sci. Process Impacts* **16**, 1841–1859 (2014).
39. Fairbanks, R. G., Sverdlove, M., Free, R., Wiebe, P. H. & Bé, A. W. H. Vertical distribution and isotopic fractionation of living planktonic foraminifera in the Panama Basin. *Nature* **298**, 841–844 (1982).
40. Ravelo, A. C. & Fairbanks, R. G. Oxygen isotopic composition of multiple species of planktonic foraminifera: recorders of modern photic zone temperature gradient. *Paleoceanography* **7**, 815–831 (1992).
41. Farmer, E. C., Kaplan, A., de Menocal, P. B. & Lynch-Stieglitz, J. Corroborating ecological depth preferences of planktonic foraminifera in the tropical Atlantic with the stable isotope ratios of core top specimens. *Paleoceanography* **22**, (2007).
42. Barker, S., Greaves, M. & Elderfield, H. A study of cleaning procedures used for foraminiferal Mg/Ca paleothermometry. *Geochem. Geophys. Geosyst.* **4**, 8407 (2003).
43. Altabet, M. A. et al. The nitrogen isotope biogeochemistry of sinking particles from the margin of the Eastern North Pacific. *Deep Sea Res. Part I* **46**, 655–679 (1999).
44. Deutsch, C., Sigman, D. M., Thunell, R. C., Meckler, A. N. & Haug, G. H. Isotopic constraints on glacial/interglacial changes in the oceanic nitrogen budget. *Global. Biogechem. Cycles* **4**, 1–22 (2004).
45. Farrell, J. W., Pedersen, T. F., Calvert, S. E. & Nielsen, B. Glacial–interglacial changes in nutrient utilization in the equatorial Pacific Ocean. *Nature* **377**, 514–517 (1995).
46. Devol, A. H. in *Nitrogen in the Marine Environment* 2nd edn (eds Capone, D. G. et al.) 263–301 (Academic, Burlington, 2008).
47. Adkins, J. F., McIntyre, K. & Schrag, D. P. The salinity, temperature, and $\delta^{18}O$ of the glacial deep ocean. *Science* **298**, 1769–1773 (2002).
48. Debelius, B., Gómez-Parra, A. & Forja, J. M. Oxygen solubility in evaporated seawater as a function of temperature and salinity. *Hydrobiologia* **632**, 157–165 (2009).
49. Robinson, R. S., Martinez, P., Pena, L. D. & Cacho, I. Nitrogen isotope evidence for deglacial changes in nutrient supply in the eastern equatorial Pacific. *Paleoceanography* **24**, PA4213 (2009).
50. Rafter, P. A. & Charles, C. D. Pleistocene equatorial Pacific dynamics inferred from the zonal asymmetry in sedimentary nitrogen isotopes. *Paleoceanography* **27**, PA3102 (2012).
51. Boyer, T. P. et al. *World Ocean Database 2013* (ed. Levitus, S.) (NOAA Atlas NESDIS 75, 2013).

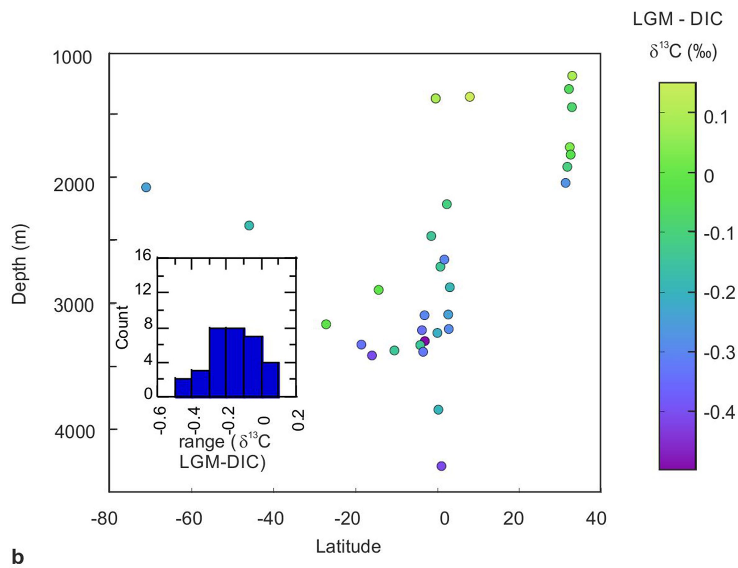


Extended Data Fig. 1 | Details of age models for ODP sites 1242 and 849. a, Matching the ODP site 1242 benthic composite $\delta^{18}O$ record to the Pacific Intermediate water stacked $\delta^{18}O$ record of ref. [29]. **b**, Matching the ODP site 849 benthic composite $\delta^{18}O$ record to the Pacific deep water stacked $\delta^{18}O$ record of ref. [29].



Extended Data Fig. 2 | Regional bulk sedimentary $\delta^{15}N$ records. Dark green, bulk sedimentary $\delta^{15}N$ record of ODP site 1242[49]; light green, bulk sedimentary $\delta^{15}N$ record of TR163-25 (this work); black, bulk sedimentary $\delta^{15}N$ record of ODP site 849[50].





Extended Data Fig. 3 | Overview and LGM evolution of carbon isotopes and oxygen concentrations in the eastern tropical Pacific. a, Dissolved oxygen concentrations (modern: North Atlantic north of 50° N, dark blue; South Atlantic south of 50° S, light blue; southeast Pacific south of 50° S, black; southwest Pacific south of 50° S, grey; northeast Pacific north of 50° N, dark purple; northwest Pacific north of 50° N, light purple; and reconstructed for the past 40 kyr: ODP site 1242, dark green; TR163-25, light green) plotted against carbon isotopes of DIC of seawater (‰) (data from refs [28,51] using https://www.nodc.noaa.gov/OC5/SELECT/dbsearch/dbsearch.html. Square boxes represent modern values at the two sites; diamonds represent LGM values (average 18–22 kyr BP). **b**, Latitudinal profile of the difference in Pacific carbon isotopes between the LGM (18–22 kyr, from epifaunal benthic foraminifera) and recent (DIC) seawater carbon isotopes (extrapolated from ref. [34]). Inset, histogram of LGM-DIC $\delta^{13}C$ (waters deeper than 1.3 km) has a normal distribution (0.1‰ bin width).

Extended Data Table 1 | Age control points for ODP sites 1242 and 849

Depth (cm) 1242	Age (ka BP) 1242		Depth (cm) 849	Age (ka BP) 849
0	0		13	6
191	11		37	10
196	13		61	17.5
255	17			
461	30.5			
565	38			

Based on matching the benthic foraminiferal composite oxygen isotope records with the stacked records of ref. [29].

Extended Data Table 2 | Age control points for TR163-25

TR163-25 depth (cm)	14C age (14C years)	Error ±1σ	ΔR	Species
40	7335	20	147±13	*G. ruber*
80	12895	45	1250±133	*N. dutertrei*
100	14250	60	1430±123	*N. dutertrei*
145	20850	130	2032±201	*N. dutertrei*

Based on ^{14}C dates and calculated reservoir ages.

LETTER

https://doi.org/10.1038/s41586-018-0604-2

A separated vortex ring underlies the flight of the dandelion

Cathal Cummins[1,2,3], Madeleine Seale[2,3,4], Alice Macente[2,4,5], Daniele Certini[1], Enrico Mastropaolo[4], Ignazio Maria Viola[1]* & Naomi Nakayama[2,3,6]*

Wind-dispersed plants have evolved ingenious ways to lift their seeds[1,2]. The common dandelion uses a bundle of drag-enhancing bristles (the pappus) that helps to keep their seeds aloft. This passive flight mechanism is highly effective, enabling seed dispersal over formidable distances[3,4]; however, the physics underpinning pappus-mediated flight remains unresolved. Here we visualized the flow around dandelion seeds, uncovering an extraordinary type of vortex. This vortex is a ring of recirculating fluid, which is detached owing to the flow passing through the pappus. We hypothesized that the circular disk-like geometry and the porosity of the pappus are the key design features that enable the formation of the separated vortex ring. The porosity gradient was surveyed using microfabricated disks, and a disk with a similar porosity was found to be able to recapitulate the flow behaviour of the pappus. The porosity of the dandelion pappus appears to be tuned precisely to stabilize the vortex, while maximizing aerodynamic loading and minimizing material requirements. The discovery of the separated vortex ring provides evidence of the existence of a new class of fluid behaviour around fluid-immersed bodies that may underlie locomotion, weight reduction and particle retention in biological and manmade structures.

Dandelions (*Taraxacum officinale* agg.) are highly successful perennial herbs that can be found in temperate zones all over the world[5]. Dandelions, as with many other members of the Asteraceae family, disperse their bristly seeds using the wind and convective updrafts[6,7]. Most dandelion seeds probably land within 2 m[8,9]; however, in warmer, drier and windier conditions, some may fly further (up to 20,000 seeds per hectare travelling more than 1 km by one estimate)[6,10]. Asteraceae seeds routinely disperse over 30 km and occasionally even 150 km[3,4].

Plumed seeds comprise a major class of dispersal strategies used by numerous and diverse groups of flowering plants, of which the common dandelion is a representative example. Plumed seeds contain a bundle of bristly filaments, called a pappus, which are presumed to function in drag enhancement (Fig. 1a–c). The pappus prolongs the descent of the seed, so that it may be carried further by horizontal winds[11], and may also serve to orientate the seed as it falls[7,12].

Dandelion seeds fall stably at a constant speed in quiescent conditions[2,13–15]. For wind-dispersed seeds, maintaining stability while maximizing descent time in turbulent winds may be useful for long-distance dispersal[16,17]. It is not clear, however, why plumed seeds have opted for a bristly pappus rather than a wing-like membrane, which is known to enhance lift in some other species (for example, maples[1]). Here we analyse the flight mechanism of the dandelion by characterizing the fluid dynamics of the pappus and identifying the key structural features enabling its stable flight.

To examine the flow behaviour around the pappus, we built a vertical wind tunnel (Fig. 1d and Methods), which was designed so that the seed can hover at a fixed height. The flow past the pappus was visualized for both freely flying (Supplementary Video 1) and fixed (Fig. 1e, f and Supplementary Videos 2, 3) samples, using long-exposure photography and high-speed imaging. We found a stable air bubble (a vortex ring) that is detached from the body, yet steadily remains a fixed distance downstream of the pappus (Fig. 1e, f and Extended Data Figs. 1a–j, 2a–j, 3a–d). Bluff bodies (such as circular disks) may generate vortex rings in their wake, but these are either attached to the body or shed from it and advected downstream. The vortex ring in the wake of the pappus is neither attached nor advected downstream, and we therefore called this vortex a separated vortex ring (SVR). The topology of SVRs has been considered theoretically, but was thought to be too unstable to actually occur[18]; here we show that the design of the pappus stabilizes the SVR.

Attached vortex rings form behind circular obstacles; however, it is unclear how the pappus can generate a vortex ring with such a limited air–structure interface (that is, high porosity). The morphology of the dandelion seeds was determined using X-ray computed microtomography (μCT) and light microscopy (Fig. 1a–c and Methods). The pappus was found to comprise $n = 100$ filaments (95–106 (mean (95% confidence interval)); $n = 10$ seeds) that radiate out from a central point (the pulvinus), each with a mean length (L) of 7.4 mm (7.35–7.46 mm (95% confidence interval); $n = 937$ filaments; Fig. 1a, b) and mean diameter (d) of 16.3 μm (15.7–17.0 μm (95% confidence interval); $n = 10$ filaments; Fig. 1c). The porosity (ε, defined as the ratio of the empty projected area to the plan area of the enclosing disk) of the pappus was measured using light microscopy (Methods) and was found to be 0.916 (0.907–0.923 (mean (95% confidence interval); $n = 10$ seeds).

The Reynolds number is a non-dimensional parameter characterizing the relative importance of inertial to viscous forces in a fluid. The flow through and around the pappus involves two different Reynolds numbers: that of the entire pappus ($\mathrm{Re} = UD/\nu$, in which U is the velocity of the seed, D is the diameter of the pappus and ν the kinematic viscosity of the fluid) and that of an individual filament ($\mathrm{Re_f} = Ud/\nu$). Our modelling revealed that the pappus of a dandelion benefits from a ‘wall effect’[19,20] at low $\mathrm{Re_f}$ (Methods). Neighbouring filaments interact strongly with one another because of the thick boundary layer around each filament, which causes a considerable reduction in air flow through the pappus (Methods). This effect—which was previously considered to be unimportant for dandelion seeds[2,21]—confers the high drag coefficient of the seed, which helps the seed to remain aloft.

The drag coefficient ($C_D = F/0.5\rho U^2 A$, in which F is the drag force acting on the seed, ρ is the density of air and A is the projected area of the pappus) of the dandelion seeds was calculated by measuring the terminal velocity $U = 39.1$ cm s^{-1} (34.9–43 cm s^{-1}; mean (95% confidence interval); $n = 10$ seeds) in a drop test (Fig. 2a). The seeds were ballasted and cut to vary the weight to explore a wide range of Re (Methods). The mean diameter of the dandelion pappi in our drop tests was $D = 13.8$ mm (13.2–14.3 mm (95% confidence interval); $n = 10$ seeds). With a mean porosity of $\varepsilon = 0.916$, the total projected area of

[1]School of Engineering, Institute for Energy Systems, University of Edinburgh, Edinburgh, UK. [2]School of Biological Sciences, Institute of Molecular Plant Sciences, University of Edinburgh, Edinburgh, UK. [3]SynthSys Centre for Systems and Synthetic Biology, University of Edinburgh, Edinburgh, UK. [4]School of Engineering, Institute for Integrated Micro and Nano Systems, University of Edinburgh, Edinburgh, UK. [5]School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK. [6]Centre for Science at Extreme Conditions, University of Edinburgh, Edinburgh, UK. *e-mail: i.m.viola@ed.ac.uk; naomi.nakayama@ed.ac.uk

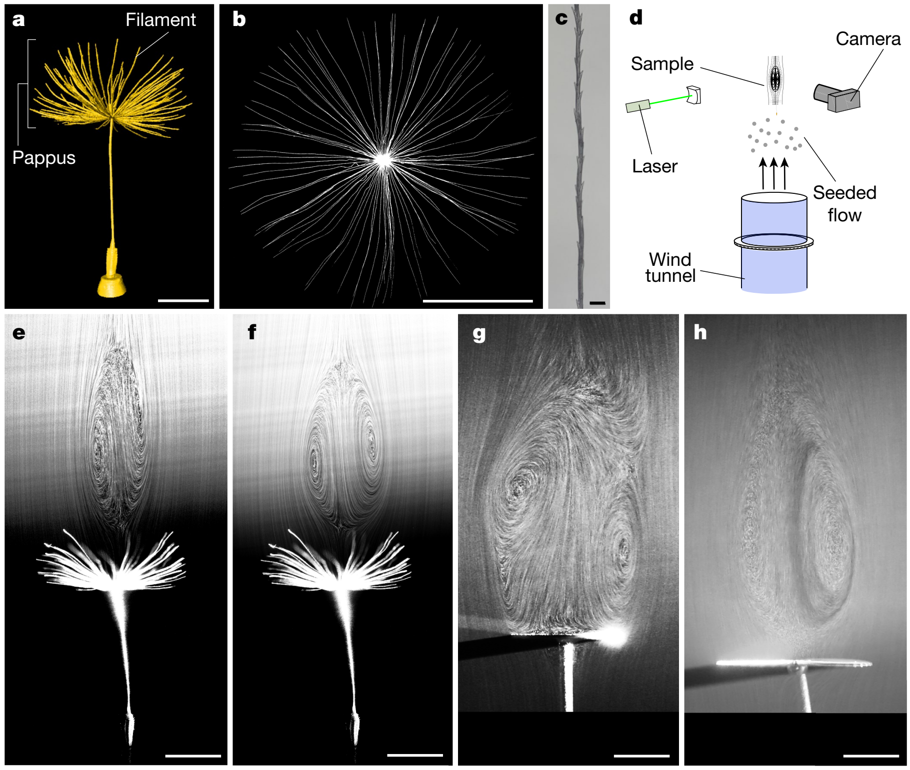


Fig. 1 | The dandelion seed and the vortex that it generates. **a**–**c**, Structural features of the drag-generating pappus at multiple scales: the μCT scan of a dandelion seed (**a**), the top-down view of the pappus (**b**) and the light microscopy image of a section of a filament (**c**). **d**, **e**, A vertical wind tunnel (**d**) was used to visualize the steady vortex downstream of a dandelion seed (**e**) at the terminal velocity of a seed. **f**, At 60% of the terminal velocity, the vortex is slightly larger and more symmetric, showing the structure of the separated vortex ring more clearly. **g**, **h**, In the same flow conditions as **e** and **f**, solid and porous disks generate vortex shedding (**g**) and a separated vortex ring (**h**), respectively. Scale bars, 50 μm (**c**) or 5 mm (all other panels).

the pappus is $A = 12.6\ \text{mm}^2$ (11.5–$13.5\ \text{mm}^2$). For a solid disk to supply the same drag force (that is, with a mean weight (W) of $6.2\ \mu\text{N}$ (5.51–$6.86\ \mu\text{N}$ (95% confidence interval); $n = 10$ seeds) of the seed) as the pappus at the same terminal velocity (see Methods), its diameter is given by $D_{\text{disk}} = \sqrt{8W/(1.17\rho\pi U^2)} = 8.6$ mm, which is 38% smaller than D. The Re of the pappus is 357, whereas the equivalent disk has an Re of 222.

The ratios of the equivalent disk diameter (D_{disk}) and area (A_{disk}) to the pappus diameter (D) and area (A), respectively, indicate that the equivalent disk is always smaller, but has a significantly higher projected area than the pappus (Fig. 2b). Thus, the pappus delivers more than four times the amount of drag per unit area compared to a solid disk[22], which quadruples C_D. The pappus achieves this effect through the interaction between the thick boundary layers surrounding each filament (Methods). In terms of material requirement, the pappus has a volume of less than $77.5\ \text{pm}^3$ (given that individual filaments are more than 50% hollow[15]). An equivalent impervious membrane of this volume would be about 1 μm in thickness, which is far thinner than the wings of flying seeds[14], although the composition of the material may also affect the efficiency of construction.

The existence of the SVR and the elevated drag coefficient are a consequence of the filaments considerably reducing the flow through the pappus, decreasing its permeability. In turn, the pressure downstream of the pappus is reduced, which increases the drag on the pappus (Methods). We measured the flow using particle image velocimetry (PIV, Methods) in the vortex downstream of the dandelion pappus (Fig. 3a–c); the magnitude of the maximum reverse flow was about 10% of the freestream velocity. We distinguished between attached and separated stable vortex rings based on the position of the upstream stagnation point (z_{su}): if $z_{\text{su}} > 0$, the vortex is separated (see Fig. 3a); otherwise it is attached.

To explore the effects of porosity, silicon disks mimicking the pappus were microfabricated, for which the degree of the porosity varied from 0 (that is, impervious) to 0.92 (comparable to a pappus) (Methods and Extended Data Fig. 4a–p). The disks were held fixed in position in the vertical wind tunnel, and flow visualization was used to explore the flow dynamics across the same range of Re as for our biological samples (for example, Fig. 1g, h). All disks generated a prominent recirculating wake (Fig. 3d–f and Supplementary Videos 4, 5). As ε increases, this vortex detaches from the disk to form an SVR. The structure and nature of the vortex depends on Re and ε. For low Re, the vortex is axisymmetric, but it loses this symmetry as Re increases. This was also observed on the dandelion pappus (Fig. 1e, f).

Our PIV analysis revealed that the magnitude of the maximum reverse flow for disks is on the order of 10% of the freestream velocity, which is in good agreement with our analysis of the flow around the biological samples (Fig. 3a–c). In both the disks and the biological samples, the streamwise length of the SVR is equal to about one characteristic diameter (disk and pappus diameter, respectively).

The SVR is not always steady; for a given porosity ε, there is a critical Reynolds number (Re_c) at which the SVR breaks down into periodic vortex shedding (Extended Data Fig. 5a–l and Supplementary Discussion). We measured Re_c for the dandelion seeds and porous disks (Methods). For the impervious disk, the measured Re_c (149 ± 2, combined the mean ± s.e.m. of velocity, diameter and kinematic viscosity measurements) is consistent with existing results of direct numerical simulations[23,24], therefore validating our experimental methodology.

Identification of Re_c for the disks and dandelion samples (Fig. 4a) revealed that Re_c generally increases with increasing ε. Figure 4a, b shows the boundary in the Re–ε parameter space that separates regions of steady SVRs and unsteady vortex shedding for porous disks. The mean measured Re_c for dandelion seeds was $\text{Re}_c = 429$ (415–440 (95% confidence interval); $n = 10$), which is in good agreement with the Re_c

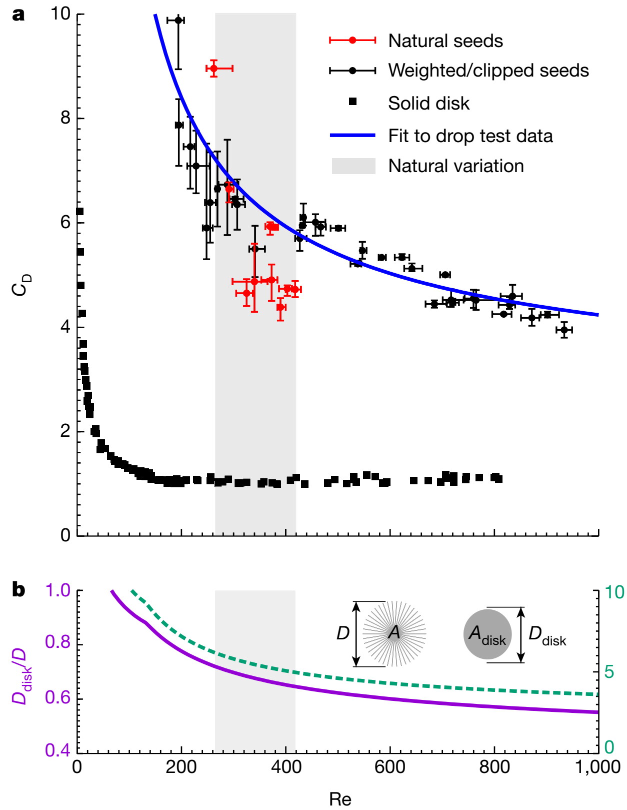


Fig. 2 | The forces on dandelion seeds compared with those on solid disks. a, The drag coefficient C_D for natural (red filled circle) and artificially weighted/clipped (black filled circle) dandelion seeds as a function of Re. The experimental data are a pool of $n = 10$ independent biological samples dropped a total of 55 times. In each of the 55 drops, the weight and velocity was measured multiple times, and the error bars are mean and 95% confidence intervals. The blue curve indicates the fit to all of the drop test data. The C_D for a solid disk from previous experiments is also shown (filled square)[22]. **b**, The ratios of the equivalent disk diameter to pappus diameter D_{disk}/D (solid magenta curve) and equivalent disk area to pappus area A_{disk}/A (dashed green curve; see insets), showing the dimensions of the impervious disk that generates the same drag as the pappus at the same velocity. The curves plotted in **b** are obtained from fitting to the data in **a**. **a**, **b**, The shaded area spans the range of the biological variation in Re for dandelion seeds.

value of 457 ± 5 (combined mean ± s.e.m. of velocity, diameter and kinematic viscosity measurements) that was found for a porous disk with identical porosity (see Methods). This result indicates that, despite its geometric complexity spreading in height, the pappus acts as if it is a flat circular disk with the same porosity.

All of the dandelion samples that we tested flew at a Re below Re_c (Fig. 4b). This suggests that evolution has tuned the pappus porosity to eliminate vortex shedding as the seed flies. Therefore, the filamentous design of plumed seeds confers two major advantages compared to a membranous one: a fourfold increase in the loading and an enhancement of the flight stability. This makes the plumed design far more efficient at flying than a membrane (that is, a circular disk) for light-weight seeds.

Traditional mathematical models of the pappus of a dandelion seed rely on the assumption that each of the filaments of the seed can be treated as a translating cylinder, with the total drag on the pappus being the sum of the contributions from each filament[2,14]. However, our numerical modelling and experimental measurements revealed that the flow through the seed entails strong interactions between neighbouring filaments[19,20], causing the pappus to behave as a permeable membrane. It has been suggested that changing the permeability of a body could be useful to control or suppress the vortex shedding[25]. A recent study has confirmed that the motion of freely falling disks (with $Re > 10^3$) can be stabilized by a hole in the centre of the disk[26]. An oscillating wake is a necessary contributing factor for the unsteady motion of falling disks[26], and the dandelion seed has eliminated this oscillation by evolving a pappus with a high porosity, thus enabling steady flight.

The initial motion of dandelion seeds is brief but fast, and is rapidly stabilized[15] into an equilibrium orientation that minimizes the terminal velocity of the seed, allowing the seed to make maximal use of updrafts[17]. Our experiments demonstrate that the stabilization of plumed seeds is not guaranteed by an arbitrarily porous pappus, as was previously suggested[11,12]. Instead, stability is gained by tuning the porosity of the pappus.

There are two major types of wind-dispersed seeds, which are distinguished by their appendage (winged or plumed) or equally by their flight mechanism (lift- or drag-based, respectively)[14]. The preferred mode of flight for large seeds—such as the maple seed—is winged[1,2], where high lift forces are attained by a leading-edge vortex. The leading-edge vortex reduces the pressure on the upper face of the wing, enhancing lift compared with non-rotating winged seeds. For winged seeds, greater release heights are necessary to reach the stable lift-generating phase. Therefore, winged flight is probably not effective for the dispersal of small and light seeds of short plants. Instead, the bristly pappus

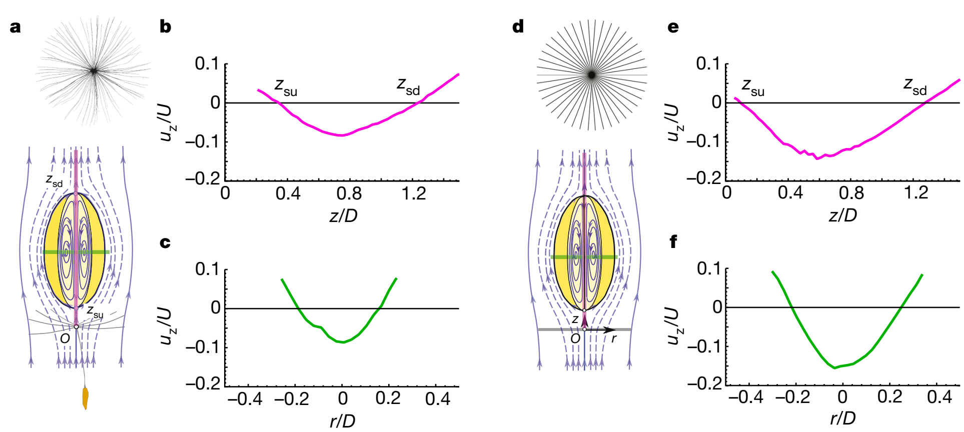


Fig. 3 | Flow diagnostics of the SVR for the dandelion seeds and a circular disk with comparable porosity. a–**c**, Dandelion pappus. Data are representative of $n = 10$ biological replicates. **d**–**f**, A circular disk with comparable porosity. **a**, Schematic view of the SVR (yellow) and the streamlines (blue) past the pappus (bottom) and a plan view of the pappus (top). The origin of the coordinate system $O(r,z)$ is the centre of the base of the filaments with the streamwise coordinate z pointing downstream and the radial coordinate r. **b**, **c**, The axial velocity u_z was measured along the z (magenta) and r (green) directions with PIV (**b** and **c**, respectively). Note that u_z is non-dimensionalized with U, whereas z and r are non-dimensionalized with D. **d**–**f**, The same as **a**–**c** for a disk with a porosity of 0.89 and a computer-aided design drawing of the porous disk in plan view (**d**, top). **e**, **f**, Data were obtained for a single disk.

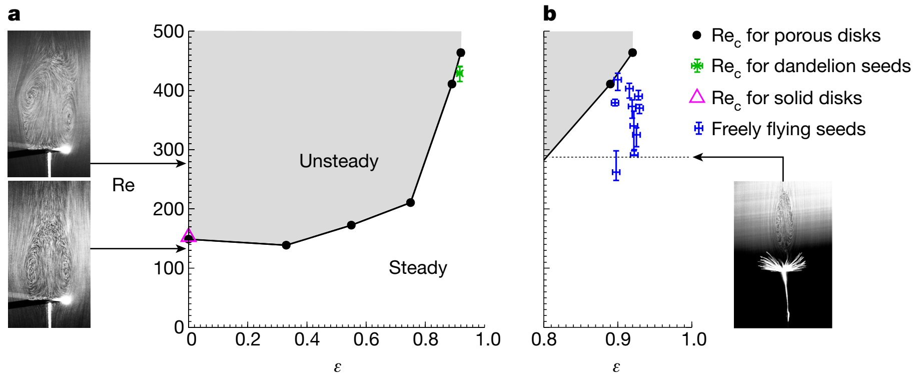


Fig. 4 | The loss of stability of the wakes past porous disks and dandelion seeds. a, The limiting Reynolds number at which the wake becomes unstable (Re_c) is plotted on the Re–ε plane for porous silicon disks. The mean Re_c for the dandelion samples (green), and values from the literature[23] for solid disks (magenta). The green data point shows the mean and 95% confidence intervals, $n = 10$ independent biological repeats. Each black dot is obtained from data from a single disk at the stated porosity. **b**, A zoomed-in region for $0.8 < \varepsilon < 1$, on which the measured values of Re and ε for freely flying dandelion seeds (blue) are superimposed. Blue data are mean and 95% confidence intervals, $n = 10$ independent biological repeats. Data for Reynolds numbers are identical to those in Fig. 2a. Insets show snapshots of the flow at the indicated Re behind solid disks (left side) and dandelion pappus (right side).

of the dandelion enhances its flight capacity through drag using a completely different type of vortex.

The shift from membranous to bristle-based flight occurs in animals, too: very small insects (for example, *Thrips physapus* L.) have evolved bristly wings rather than membranous ones[20,27–29]. Flight at this scale makes use of a technique called 'clap and fling', and bristly wings reduce the force required to fling the wings apart[28,30]. These insects can also float by spreading out their wings, generating 90% of the wing loading of a solid plate with 10% of the material[20]. Bristly appendages are common among light-weight fliers and swimmers, and it is likely that the SVR and similar permeability-dependent vortices have a crucial role in their locomotion. They may also underlie the feeding mechanisms of underwater organisms, such as the larvae of the black fly (*Simulium vittatum*), which use a bristly fan for suspension feeding[31,32]. Because Re_c shifts with the degree of porosity, small changes in the morphology of their appendages may markedly affect the dynamics of this vortex, leading to a switch in their biological function, for example, from foraging to escape[33].

By uncovering the physics behind the flight of the dandelion, we have discovered a novel type of fluid behaviour around fluid-immersed bodies. As filamentous microstructures within the relevant Re regimes (<1 for the pore scale and about 100–1,000 for the body scale) are commonplace in the biological world[19,31,34], we anticipate that permeability-dependent flow control is prevalent in nature. Traditionally, fluid dynamics investigations tend to observe a single Re scale; exploration of interactions among multiple Re regimes may uncover other as yet unknown fluid behaviours.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0604-2.

Received: 19 April 2018; Accepted: 21 August 2018;
Published online 17 October 2018.

1. Lentink, D., Dickson, W. B., van Leeuwen, J. L. & Dickinson, M. H. Leading-edge vortices elevate lift of autorotating plant seeds. *Science* **324**, 1438–1440 (2009).
2. Greene, D. F. & Johnson, E. A. The aerodynamics of plumed seeds. *Funct. Ecol*. **4**, 117–125 (1990).
3. Ridley, H. N. On the dispersal of seeds by wind. *Ann. Bot*. **os-19**, 351–364 (1905).
4. Small, J. The origin and development of the Compositæ. *New Phytol*. **17**, 200–230 (1918).
5. Holm, L. G. *World Weeds: Natural Histories and Distribution* (John Wiley & Sons, New York, 1997).
6. Tackenberg, O., Poschlod, P. & Kahmen, S. Dandelion seed dispersal: the horizontal wind speed does not matter for long-distance dispersal—it is updraft! *Plant Biol*. **5**, 451–454 (2003).
7. Sheldon, J. & Burrows, F. The dispersal effectiveness of the achene–pappus units of selected Compositae in steady winds with convection. *New Phytol*. **72**, 665–675 (1973).
8. Nathan, R. et al. Mechanisms of long-distance seed dispersal. *Trends Ecol. Evol*. **23**, 638–647 (2008).
9. Soons, M. B. & Ozinga, W. A. How important is long-distance seed dispersal for the regional survival of plant species? *Divers. Distrib*. **11**, 165–172 (2005).
10. Greene, D. F. The role of abscission in long-distance seed dispersal by the wind. *Ecology* **86**, 3105–3110 (2005).
11. Andersen, M. C. An analysis of variability in seed settling velocities of several wind-dispersed Asteraceae. *Am. J. Bot*. **79**, 1087–1091 (1992).
12. Burrows, F. Calculation of the primary trajectories of plumed seeds in steady winds with variable convection. *New Phytol*. **72**, 647–664 (1973).
13. Andersen, M. C. Diaspore morphology and seed dispersal in several wind-dispersed Asteraceae. *Am. J. Bot*. **80**, 487–492 (1993).
14. Minami, S. & Azuma, A. Various flying modes of wind-dispersal seeds. *J. Theor. Biol*. **225**, 1–14 (2003).
15. Sudo, S., Matsui, N., Tsuyuki, K. & Yano, T. Morphological design of dandelion. In *Proc. 11th International Congress and Exposition* (Society for Experimental Mechanics, 2008).
16. Tackenberg, O., Poschlod, P. & Bonn, S. Assessment of wind dispersal potential in plant species. *Ecol. Monogr*. **73**, 191–205 (2003).
17. Stevenson, R. A., Evangelista, D. & Looy, C. V. When conifers took flight: a biomechanical evaluation of an imperfect evolutionary takeoff. *Paleobiology* **41**, 205–225 (2015).
18. Délery, J. *Three-Dimensional Separated Flows Topology: Singular Points, Beam Splitters and Vortex Structures* (John Wiley & Sons, 2013).
19. Vogel, S. *Life in Moving Fluids: The Physical Biology of Flow* (Princeton Univ. Press, Princeton, 1981).
20. Barta, E. & Weihs, D. Creeping flow around a finite row of slender bodies in close proximity. *J. Fluid Mech*. **551**, 1–17 (2006).
21. Casseau, V., De Croon, G., Izzo, D. & Pandolfi, C. Morphologic and aerodynamic considerations regarding the plumed seeds of *Tragopogon pratensis* and their implications for seed dispersal. *PLoS ONE* **10**, e0125040 (2015).
22. Roos, F. W. & Willmarth, W. W. Some experimental results on sphere and disk drag. *AIAA J*. **9**, 285–291 (1971).
23. Shenoy, A. & Kleinstreuer, C. Flow over a thin circular disk at low to moderate Reynolds numbers. *J. Fluid Mech*. **605**, 253–262 (2008).
24. Fernandes, P. C., Risso, F., Ern, P. & Magnaudet, J. Oscillatory motion and wake instability of freely rising axisymmetric bodies. *J. Fluid Mech*. **573**, 479–502 (2007).
25. Cummins, C., Viola, I. M., Mastropaolo, E. & Nakayama, N. The effect of permeability on the flow past permeable disks at low Reynolds numbers. *Phys. Fluids* **29**, 097103 (2017).
26. Vincent, L., Shambaugh, W. S. & Kanso, E. Holes stabilize freely falling coins. *J. Fluid Mech*. **801**, 250–259 (2016).
27. Davidi, G. & Weihs, D. Flow around a comb wing in low-Reynolds-number flow. *AIAA J*. **50**, 249–253 (2012).
28. Jones, S. K., Yun, Y. J. J., Hedrick, T. L., Griffith, B. E. & Miller, L. A. Bristles reduce the force required to 'fling' wings apart in the smallest insects. *J. Exp. Biol*. **219**, 3759–3772 (2016).
29. Lee, S. H. & Kim, D. Aerodynamics of a translating comb-like plate inspired by a fairyfly wing. *Phys. Fluids* **29**, 081902 (2017).
30. Santhanakrishnan, A. et al. Clap and fling mechanism with interacting porous wings in tiny insect flight. *J. Exp. Biol*. **217**, 3898–3909 (2014).
31. Cheer, A. & Koehl, M. Paddles and rakes: fluid flow through bristled appendages of small organisms. *J. Theor. Biol*. **129**, 17–39 (1987).
32. Ross, D. H. & Craig, D. A. Mechanisms of fine particle capture by larval black flies (Diptera: Simuliidae). *Can. J. Zool*. **58**, 1186–1192 (1980).

33. van Duren, L. A. & Videler, J. J. Escape from viscosity: the kinematics and hydrodynamics of copepod foraging and escape swimming. *J. Exp. Biol.* **206**, 269–279 (2003).
34. Seale, M., Cummins, C., Viola, I. M., Mastropaolo, E. & Nakayama, N. Design principles of hair-like structures as biological machines. *J. R. Soc. Interface* **15**, 20180206 (2018).

Acknowledgements This work was supported by the Leverhulme Trust (RPG-2015-255) and the Royal Society (UF140640). We thank I. Butler (Geosciences, University of Edinburgh) for assistance with the μCT scans; and A. Firth and M. Mason (Engineering, University of Edinburgh) for helping to build the wind tunnel.

Reviewer information *Nature* thanks M. Dickinson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.C., E.M., I.M.V. and N.N. designed the experiments. C.C. designed and set up the wind tunnel. C.C. carried out the numerical analyses, the flight assay and flow visualization with assistance from M.S. and D.C. C.C. designed and E.M. fabricated the silicon disks. A.M. optimized and performed the μCT scans, and M.S. analysed the resulting 3D images. C.C. wrote the manuscript; M.S., E.M., I.M.V. and N.N. helped with revision and editing. E.M., I.M.V. and N.N. designed and oversaw the project; I.M.V. supervised the investigations of fluid mechanics and N.N. supervised the biological and structural studies.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0604-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0604-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to I.M.V. or N.N.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Experiments using vertical wind tunnel. *Flow visualization.* A vertical wind tunnel was built to visualize the flow around natural and artificial (microfabricated) pappi. Airflow was generated in the wind tunnel using a fan (San Ace 9GA0412P6G001) mounted on the inlet of the tunnel. The fan speed was controlled by pulse wave modulation using an Arduino Uno. The flow conditioning consisted of three meshes of open area ratios: 70%, 66% and 66% and a honeycomb flow straightener. A fog machine (Rosco 1700) was used to seed the air in the wind tunnel, and the flow was illuminated using a 2.5-W diode-pumped continuous wave laser (Medialas DPGL 2500) with a wavelength of 532 nm (Extended Data Fig. 6b).

The videos were obtained using a Canon EOS 70D (for low-speed videos) and a Fastcam Photron SA1 (for high-speed videos)—both with a Tamron 180 mm F3.5 SP AF Di Macro Lens. The flow speed and turbulent intensity ($T_u = u'/U$, in which u' is the root-mean-square of the turbulent velocity fluctuations and U is the mean velocity) was measured using two-dimensional laser Doppler anemometry, comprising a 5-W argon ion laser with a wavelength of 488 nm and Dantec Dynamics FibreFlow optics (Extended Data Fig. 6a). The mean velocity varied from a minimum of 0.19 m s^{-1} to a maximum of 0.72 m s^{-1}, with a maximum T_u of 3.6% and a mean T_u of 2.3%. The flow past $n = 10$ dandelion seeds—randomly selected from ten different capitula (see 'Growth of dandelion specimens (*T. officinale* agg.))—flying at their terminal velocity was visualized in the wind tunnel; each seed showed a prominent SVR in its wake (Supplementary Video 1).

The same ten seeds that were used for the freely flying experiments were fixed to trestles, and held fixed in the wind tunnel at different flow speeds; long-exposure photographs of the flow past these seeds were obtained using a Canon EOS 70D. Downstream of each seed, an SVR is present, which is stable across the entire biological variation of the seeds (Fig. 1e, f).

PIV was performed using the CW laser to illuminate air that is seeded with smoke, and a high-speed camera; the data were post-processed using the MATLAB toolbox PIVlab 1.41. The frame rate used in the PIV experiments varied from a minimum of 50 f.p.s. (for low flow-speed experiments) to a maximum of 250 f.p.s. (for high flow-speed experiments). We used a multi-pass linear window deformation technique: the first pass used a 64 × 64-pixel interrogation window, the second pass 32 × 32-pixel interrogation window and sub-pixel displacement was estimated using two-dimensional Gaussian regression. Because we were interested in the flow far from the body, masking was not necessary, and therefore, no masking was used.

Detection of vortex shedding. The detection of vortex shedding from dandelion samples and replica pappi in the wind tunnel experiments was measured using MATLAB to compute the structural similarity index between a reference frame of the wake region and subsequent frames of the video[35,36]. The power spectral density estimate of this signal was then found using the covariance method, and the peak frequency f was extracted for a range of Re. A non-zero f indicated the presence of vortex shedding. To compute Re_c for our dandelion seeds, we analysed the flow past $n = 10$ seeds, which were fixed in our wind tunnel. We found that the 95% confidence interval for the mean Re_c for the dandelion samples was $Re_c = 429$ (415–440) (mean (95% confidence interval); $n = 10$ seeds).

To compare the mean Re_c for dandelion samples to the value of Re_c predicted by our porous disks, we linearly interpolated between the data points in Fig. 4b. From this, we estimated that the Re_c of porous disks at the same porosity as our dandelion samples is $Re_c = 457 \pm 5$ (combined mean ± s.e.m. of velocity, diameter and kinematic viscosity measurements) (Fig. 4a, b).

Growth of dandelion specimens (*T. officinale* agg.). Dandelion seeds were collected from a single plant growing in Edinburgh (55.922684 °N, 3.170703 °W) in April 2014. Seeds were germinated in 10-cm round Petri dishes containing distilled water in 16 h light/8 h dark conditions (100 μmol m^{-2} s^{-1}, 25 °C during the day, 23 °C during the night) for two weeks. They were then transplanted to 7 × 7 × 8-cm^3 pots with soil/perlite mix 60% v/v Levington's F2+S (Everris), 24% v/v standard perlite (Sinclair), 16% v/v washed horticultural sand 0.3 g l^{-1} Exemptor (Everris)) and grown in 16 h light/8 h dark conditions in a room with a controlled environment (100 μmol m^{-2} s^{-1}, 21 °C) for four weeks. Plants were transplanted into 4-l pots with peat/sand mix (83% v/v medium peat (Clover), 21% v/v washed horticultural sand, 3 g l^{-1} garden limestone (Arthur Bowers), 1 g l^{-1} Osmocote Exact Standard 5–6 months (Everris), 0.4 g l^{-1} Exemptor (Everris)) and transferred to a glasshouse with ambient light supplemented to ensure a 16-h day (minimum intensity of 250 μmol m^{-2} s^{-1}, 06:00–22:00 GMT) and temperature of 21 °C during the day, 18 °C during the night. For μCT scans, seeds used were the offspring of the original collected plants.

For all other experiments, seeds were from the subsequent generation. All of these seeds from the second generation originated from the same parent plant. As *T. officinale* is apomictic, all seeds are assumed to be genetically identical. Throughout the paper, we use the term dandelion 'seed' to refer to the entire diaspore (fruit–pappus unit).

X-ray computed microtomography (μCT). Ten dandelion samples were individually attached to machined sharpened carbon cones using forceps and cyanoacrylate glue (RS Pro). Samples were sputter-coated with gold for 100–200 s (corresponding to a thickness of approximately 150–300 nm). Scan settings were as shown in Extended Data Table 1.

Data were reconstructed using Octopus 7 software[37]. The voxel (three-dimensional pixel) size of the reconstructed μCT datasets was 25 μm.

Post-processing of the reconstructed data was carried out with Avizo 9.0.1 (FEI, ThermoFisher Scientific) and R[38]—see Extended Data Fig. 7 for the workflow chart. Scans were filtered by unsharp masking with a three-voxel kernel size. Small holes of up to 26 voxels were filled and a labelled image was created by interactive thresholding.

For analysis of the pappus geometry, the segmented data were skeletonized using an implementation of the TEASAR algorithm[39] (scale = 2.5, constant = 4), in which a tree structure is formed from traced peaks of distance maps and looping is not permitted. The starting point for skeletonization was manually selected for each sample to begin at the central point of the pappus (the pulvinus, where all filaments are attached). Nodes with a coordination number of one (that is, connected to no more than one other node) were considered potential filament end points and nodes were visually inspected to remove false positives from further analysis. A coordinate mapping of each filament was obtained by finding the shortest path between the central starting point and each filament end.

The point coordinates along the length of each filament were smoothed using a Gaussian smoothing filter (window size = 16, $\alpha = 2.5$, tails retained). The window size was selected by stepwise increases of the window size until the mean filament arc length changed by less than 1% from its previous value (that is, interpolation between coordinates of the centre line was no longer significantly affected by noise arising from the limited voxel resolution of the scan). The coordinates of each spatial dimension were separately smoothed with the same settings. Points corresponding to a central disk (the pulvinus) onto which the filaments are attached were removed from further analysis (by removing a central sphere with a radius of 0.56–0.64 mm, depending on the sample), such that only filaments themselves were included.

The spacing between filaments in the pappus was estimated by calculating the distance of the centre line of each filament from the centre line of the nearest neighbouring filament. In total, 93.5% of filaments were correctly segmented, skeletonized and included in the analysis. Spacing was found to linearly increase from zero at the pulvinus to 1.32 mm at the edge of the pappus. This maximum distance divided by two represents the mean distance between filaments, and was an input into the numerical model (creeping flow past an array of filaments). As a small number of filaments were not included, the nearest neighbour calculations represent a slight overestimate. Additionally, it is important to note that these spacing distances are the spacing between centerlines. Filament diameters were at the limit of the resolution of the μCT scanner, so they were not calculated from this data.

Microscopy. *Light microscopy of individual filaments.* All filaments except one were removed from each dandelion fruit. The stalk and pulvinus were stuck onto a glass slide with a small piece of modelling clay such that the single remaining filament lay flat on the slide. Images were acquired with a Nikon E600 fluorescent microscope using a 10× objective, 1-ms exposure, 0.6 gamma and 2× gain. Each field of view was imaged 1–8 times at different focal (z) planes to account for slight changes in topography. Image processing was carried out in ImageJ to calculate filament diameters[40]. Sharp composite images were obtained for each field of view by model-based deconvolution, stitched together with linear blending and converted into binary images[41,42]. Distance maps were computed and the skeletonized centre line of the filament was overlaid. Diameters were calculated from the distance map at each pixel along the centre line of the filament. The mean of the diameter values at all points along the filament was calculated to give an overall filament diameter. The error in diameter values due to binarization was ±0.80 μm based on a pixel size of 0.40 μm.

Light microscopy of the entire pappus. The porosity of $n = 10$ dandelion seeds was measured using light microscopy. First, the mean diameter of dandelion pappi (D) were measured using a Dino-Lite digital microscope; the mean diameter was found to be $D = 13.8$ mm (13.2–14.3 mm (95% confidence interval); $n = 10$ seeds). The porosity of these pappi was then measured. The images were obtained using a Nikon SMZ1500 stereomicroscope, with 1× magnification, 38.5-ms exposure, 0.6 gamma, 1.0× gain and 1.60 saturation. The pappi were placed on a glass slide covered with 5 μl of 99% ethanol and were then covered with a glass coverslip. Overlapping sections of each pappus were imaged at different positions on the focal plane to account for the entirety of the pappus. These images were stitched together with linear blending[42] in ImageJ to form the entire pappus image. The pulvinus was inscribed in a circle to find the centre of the pappus. Images, converted to

RGB colour format, were used to calculate the empty area inside the disk, applying a colour threshold. The porosity (p) of the flattened sample was obtained by calculating the ratio of empty area to the total plan area of the pappus. The porosity ε of the original sample was then calculated to be $\varepsilon = 1 - 2L(1-p)/D = 0.916$ (0.907–0.923) (mean (95% confidence interval); $n = 10$ seeds)[2].

Error analysis using different magnifications. A single filament was removed from a dandelion fruit and placed on a glass slide. The porosity of a rectangular field of view, including a section of the filament, was measured at four different magnifications. From this, the error due to the finite resolution of the equipment was estimated to be 0.54%.

Creeping flow past an array of filaments. The Reynolds number ($\mathrm{Re} = UD/\nu$) is calculated using the pappus diameter D as the characteristic length scale, and was found to be in the order of 400. Note, however, that when discussing low-Re effects, a filament Reynolds number, based on a diameter of the filament ($\mathrm{Re_f} = Ud/\nu = 0.422$) is used. Because $\mathrm{Re_f} < 1$, the equations for creeping flow apply, and may be used to investigate the flow past the pappus. Consider the low Reynolds number flow past a body: it is well-known that the velocity boundary layer attached to the body extends many body diameters into the fluid[43], influencing the flow far from it. When this flow interacts with distant boundaries, it is known as a 'wall effect'. The following estimate of when this effect can be ignored was calculated previously[19]:

$$\frac{y}{\lambda} > \frac{20}{\mathrm{Re_f}}$$

in which λ is the characteristic length scale of the body and y is the distance to the nearest boundary. In the case of the dandelion, we are considering the effect of neighbouring filaments, so $\lambda = 16.3\,\mu\mathrm{m}$, $\mathrm{Re_f} = 0.422$ and y is the mean distance between the filaments (see 'X-ray computed microtomography (μCT)'). To neglect the influence of neighbouring filaments, we can estimate that filaments should be spaced greater than 47 filament diameters apart. However, based on our μCT scan data (see 'X-ray computed microtomography (μCT)'), the mean distance between the filaments is about 41 filament diameters, therefore, the effects of neighbouring filaments cannot be ignored.

To further confirm this hypothesis, we computed the slow flow (velocity vector ($\boldsymbol{u}$) and pressure (P)) past a rectangular array of 100 filaments with a diameter and length equal to those of the dandelion seeds. The filaments within the array were separated by a distance equal to the mean distance between the filaments. We used a previously published modelling approach[20]. The creeping flow equations

$$\nabla P = \mu \nabla^2 \boldsymbol{u}$$

$$\nabla \boldsymbol{u} = 0$$

were solved in the fluid domain (μ is the dynamic viscosity of the fluid), with each filament represented by a distribution of singularities (Stokeslets with intensity α_i and doublets with intensity β_i) along its axis.

The intensities of the singularities are computed using Wolfram Mathematica 11 by solving an appropriate system of linear equations. We used 64 points that were uniformly distributed along the axis of each body. Once the equations have been solved for $\boldsymbol{u}$ and p, the drag on each member of the pappus can be computed.

The drag exerted on the ith filament can be expressed in terms of the integral of the Stokeslet intensity along the length of the filament as follows: $D_i = 8\pi\mu \int_0^L \alpha_i(s)\,ds,\ \ i = 1,\ \ldots,\ m$

In Extended Data Fig. 8d, the drag on each filament divided by the drag of a single, isolated filament[44] D_i/D_0 is plotted. We found that there is a strong interaction between filaments. On average, a filament within the pappus experiences a reduction of 84% in drag, compared to an isolated filament. This indicates that the pappus is behaving similar to a continuous surface, substantially reducing the airflow through it. The blockage effects resulting from air being pushed around the pappus are not captured by this model. Therefore, this model cannot be used to explore the resulting flow field around the pappus.

Strictly speaking, this model is valid in the limit as $\mathrm{Re_f}$ tends to zero. However, since $\mathrm{Re_f}$ is finite, some errors are introduced[31]. Here we examine the error introduced by neglecting the small but finite $\mathrm{Re_f}$ for the filaments of the dandelion. The slow flow past an array of slender bodies has previously been analysed[27] using computational fluid dynamics for a range of small to moderate $\mathrm{Re_f}$, ranging from 0.01 to 100. This parametric study found that for $\mathrm{Re_f} \le 1$ and spacing of 10 filament diameters, the flow speed between adjacent filaments is identical to the speed found in the previously published Stokes flow model[20]. The drag force computed using the previously published model[20] differed from the force computed the parametric study[27], but the trend and order of magnitude remained very similar.

Measurements of C_D. The terminal velocity (U) of $n = 10$ dandelion seeds selected randomly from different plants was measured by dropping each seed five times. A DSLR camera (Canon EOS 70D) recorded the fall at 50 f.p.s. over 1 m. The position of the seeds was tracked using MATLAB, and the terminal velocity was found using linear regression of the tracked position data.

Additional masses (strand of polyvinylesiloxane impression material) were attached to the seeds, and the terminal velocity of the composite mass was measured as described above. The mass ((m) consisting of seed + strand) was measured using a Mettler AE 240 analytical balance. To explore the terminal velocity for masses that were lower than the natural mass of the seed, a small part of the seed was cut, and the terminal velocity of this was measured as described above.

The drag coefficient (C_D) was computed using

$$C_D = \frac{mg}{0.5\rho A U^2}$$

in which $\rho = 1.204\ \mathrm{kg\ m^{-3}}$ is the density of air at normal temperature and pressure, $g = 9.81\ \mathrm{m\ s^{-2}}$ is the acceleration due to gravity and A is the total projected area of the pappus. By adding masses, the variation in C_D across a wide range of Reynolds numbers of $\mathrm{Re} = UD/\nu$, in which $\nu = 15.11 \times 10^{-6}\ \mathrm{m^2\ s^{-1}}$ is the viscosity of air and D is the pappus diameter, was explored.

The mean mass of the dandelion seeds in our experiments was 0.633 mg (0.562–0.699 mg) (mean (95% confidence interval); $n = 10$) and the seeds fell at an average speed of $U = 39.1\ \mathrm{cm\ s^{-1}}$ (34.9–43.0 $\mathrm{cm\ s^{-1}}$) (mean (95% confidence interval), $n = 10$), leading to a mean $\mathrm{Re} = 357$.

Flow field characterization of the SVR. To characterize the SVR, we performed direct numerical simulations of the flow past a permeable circular disk with aspect ratio $\chi = d/D = 0.0011$, Darcy number $\mathrm{Da} = k/D^2 = 4.7 \times 10^{-6}$ and porosity $\varepsilon = 0.916$ at $\mathrm{Re} = UD/\nu = 175$, in which k is the permeability of the disk, and U is the freestream speed (values for porosity and diameter were obtained from morphological analysis of samples—see Extended Data Table 2). We used a previously published modelling approach[25], in which we considered the steady, axisymmetric flow past the permeable disk. In the fluid domain, the steady-state Navier–Stokes equations are solved, and inside the permeable disk, the steady-state Darcy–Brinkmann equations are solved. Continuity of the velocity and pressure are enforced at the boundary between the fluid and porous domains, and the discretized system of equations is solved using COMSOL Multiphysics.

The results from our numerical modelling are shown in Extended Data Fig. 8. The flow around and through the porous disk is characterized by a marked slowdown of velocity u_z (Extended Data Fig. 8a). This is associated with a pressure increase upstream of the disk (Extended Data Fig. 8b). Across the disk, the flow velocity is conserved while the disk subtracts potential energy from the flow, resulting in a lower pressure downstream. Subsequently, the flow downstream of the disk is affected by the adverse pressure gradient between the high pressure in the far field and the low pressure in the region downstream of the disk. This pressure gradient further slows down the flow, which eventually reversed and led to the formation of a recirculation bubble due to viscous effects (Extended Data Fig. 8a, c). Further downstream, the gradual pressure recovery enables a lower pressure gradient, that is, a lower pressure force on the fluid, which therefore recovers its velocity by entrainment of momentum from the adjacent flow streams. This results in an asymptotic increase in velocity towards the far field. We quantified the numerical uncertainty (Y_{num}), which is the sum of the uncertainties due to the grid (Y_g) and the iterative convergence (Y_c) using the approach used in previous studies[45].

We found that the numerical uncertainty in the computed value of the streamwise length of the SVR was $Y_{\mathrm{num}} < 0.02\%$ for the values of Re, Da and ε considered in this study.

The results from this numerical model provide insights into the pressure field and the general flow structure around the pappus. However, there are limitations to this simple model. The assumption of axisymmetry precludes any investigation of the observed symmetry breaking of the vortex (similar symmetry breaking is observed for impervious disks[23]) or the breakdown of the SVR into vortex shedding at higher Re. In the latter case, to compute $\mathrm{Re_c}$ using this model, the assumption of time independence would also have to be relaxed.

Topology of the SVR. Topologically, the SVR is a degenerate focus with half-saddle separation (z_{su}) and reattachment (z_{sd}) points. For low Re, the vortex is axisymmetric; however, at some $\mathrm{Re_c}$, the steady SVR loses its azimuthal symmetry by a regular bifurcation as illustrated in the schematic diagram in Extended Data Fig. 3e, f. The subsequent breakdown in stability of the SVR at $\mathrm{Re} = \mathrm{Re_c}$ is likely to occur through a Hopf bifurcation[24].

Design and microfabrication of replica pappi. Replica pappi of various porosities were designed using Wolfram Mathematica: first, a rectangle with length of l mm and varying widths of w was created using the rectangle function. The rectangle was then copied and rotated around a central point 20 times using the

Mathematica function GeometricTransformation to create a replica pappus with $n = 42$ filaments. The porosity of the disk depended on the width of the filament according to

$$\varepsilon = 1 - \frac{nw\{(l-b) + b/2\}}{\pi l^2}$$

in which $b = w/(2\tan(\pi/n))$. The resulting design was exported as a vector image for use in the microfabrication process. A length $l = 10$ mm (to explore the region $Re < 170$) and $l = 14$ mm (to explore $Re > 170$) was used.

The replica pappi were manufactured using photolithography and microfabrication techniques. A 1-μm thick layer of silicon oxide (SiO_2) was grown on a 3-inch silicon wafer substrate (thickness of 380 μm). After spincoating a 7-μm thick photoresistant film on the SiO_2 layer, the dandelion designs were patterned photolithographically onto the substrate. Afterwards, the exposed SiO_2 was removed by reaction ion etching in a plasma formed of CHF_3 and Ar. At this point, the dandelion structure was etched in deep reactive ion etching (Bosch process) using the photoresist and SiO_2 layer as an etch mask. Once the wafer was etched through completely, the dandelion structure was rinsed and bonded to an artificial stem to enable testing in the vertical wind tunnel.

Statistics. Throughout the paper, the 95% confidence intervals are obtained using bias-corrected and accelerated bootstrapping. All of the morphological data obtained from our dandelion samples was shown to be normally distributed, apart from the length filament (L).

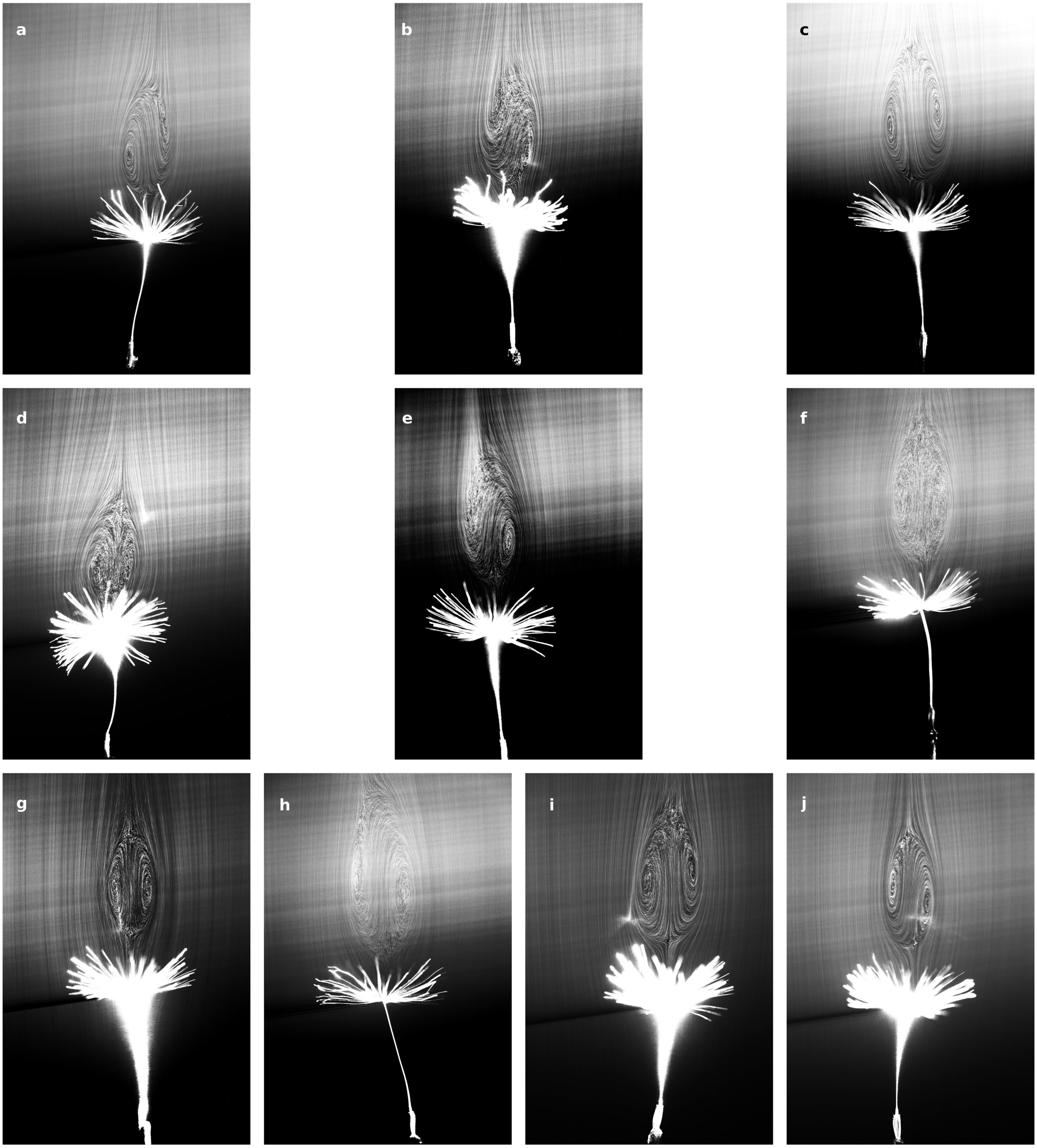
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The codes used to produce Fig. 4a are available from Edinburgh DataShare[35,36].

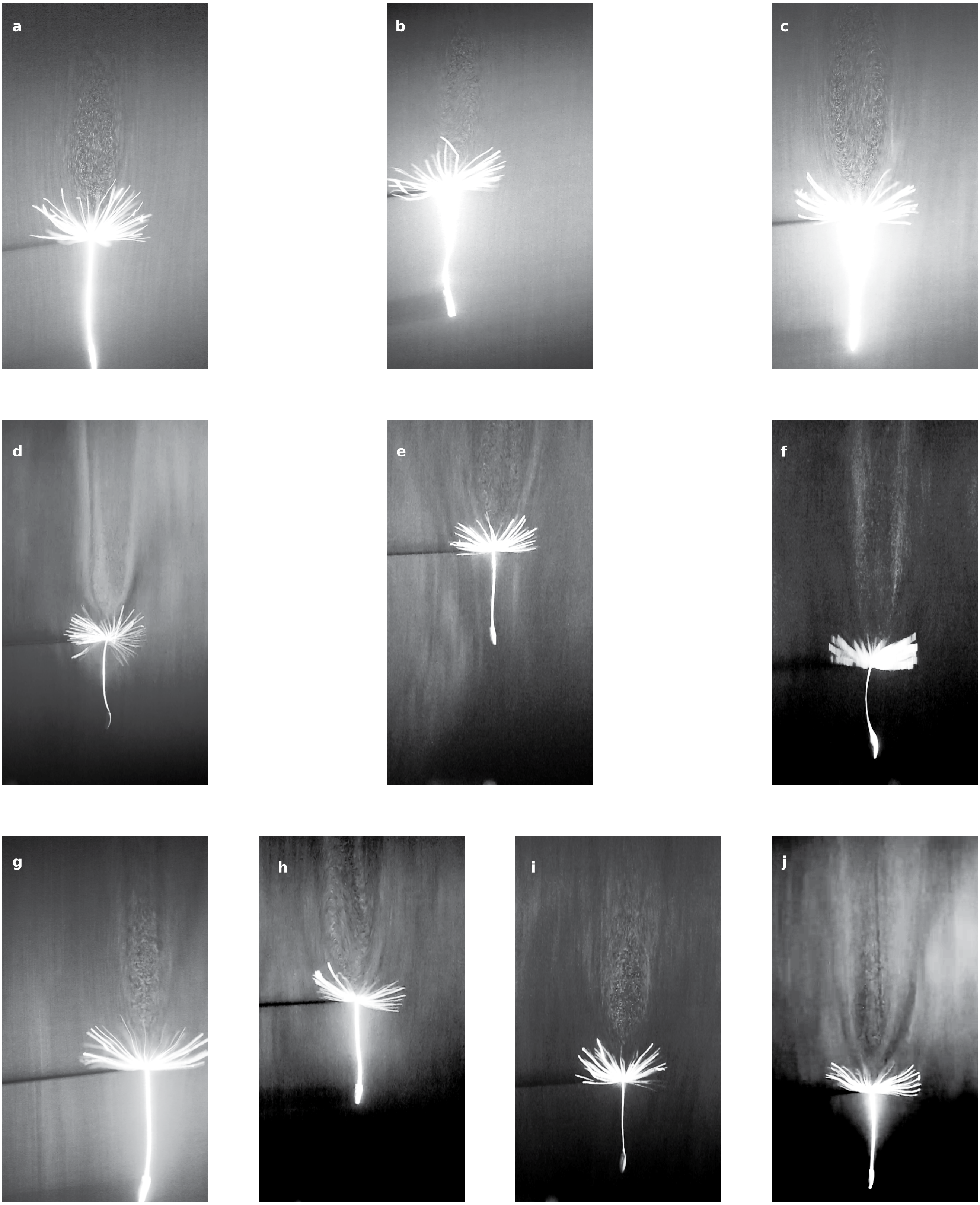
Data availability

The datasets generated and/or analysed during the current study are available from the corresponding author upon reasonable request.

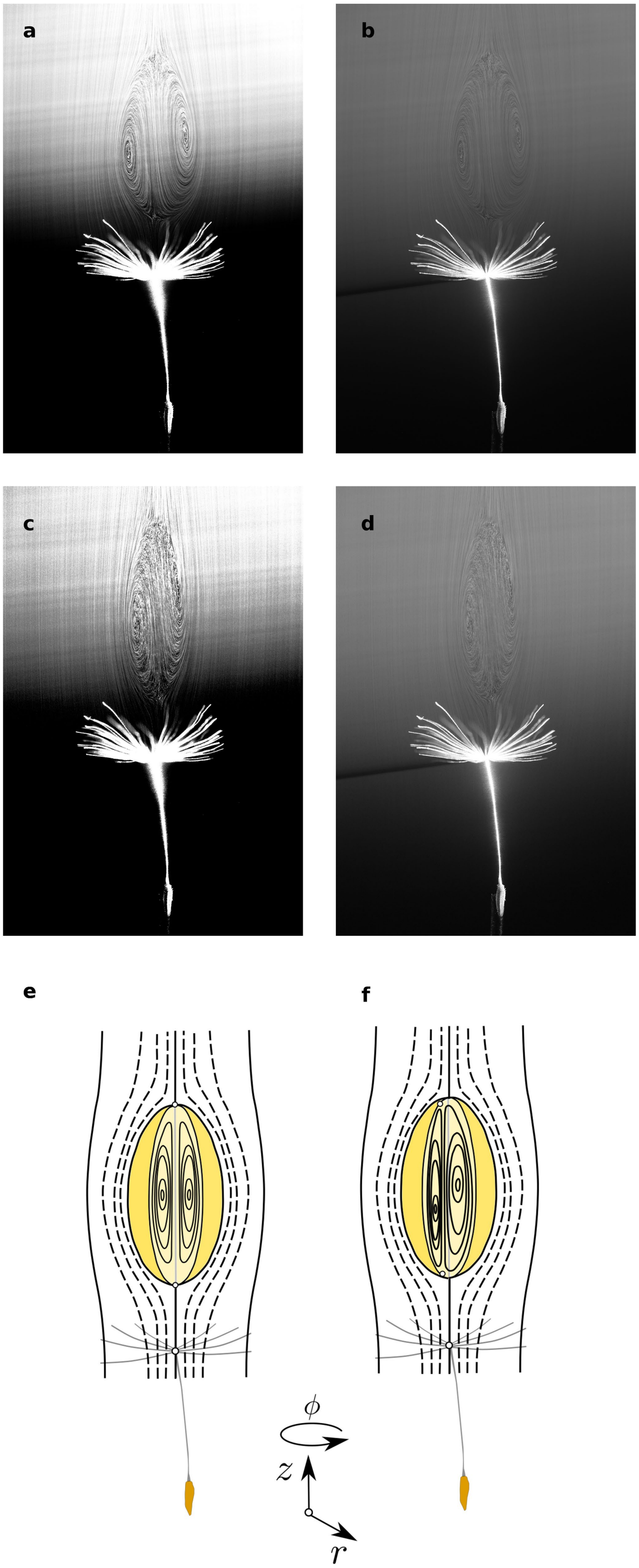
35. Cummins, C., Nakayama, N., Viola, I. M. & Mastropaolo, E. MATLAB scripts for analysis of vortex shedding. https://doi.org/10.7488/ds/2362 (2018).
36. Viola, I. M., Nakayama, N., Mastropaolo, E. & Cummins, C. Vortex shedding in the wake of a 75% porous disk. https://doi.org/10.7488/ds/2363 (2018).
37. Dierick, M., Masschaele, B. & Hoorebeke, L. V. Octopus, a fast and user-friendly tomographic reconstruction package developed in LabView®. *Meas. Sci. Technol.* **15**, 1366–1370 (2004).
38. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2013).
39. Sato, M., Bitter, I., Bender, M. A., Kaufman, A. E. & Nakajima, M. TEASAR: tree-structure extraction algorithm for accurate and robust skeletons. In *Proc. 8th Pacific Conference on Computer Graphics and Applications* (eds Barsky, B. A. et al.) 281–449 (IEEE, 2000).
40. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
41. Forster, B., Van De Ville, D., Berent, J., Sage, D. & Unser, M. Complex wavelets for extended depth-of-field: a new method for the fusion of multichannel microscopy images. *Microsc. Res. Tech.* **65**, 33–42 (2004).
42. Preibisch, S., Saalfeld, S. & Tomancak, P. Globally optimal stitching of tiled 3D microscopic image acquisitions. *Bioinformatics* **25**, 1463–1465 (2009).
43. White, C. M. The drag of cylinders in fluids at slow speeds. *Proc. R. Soc. A* **186**, 472–479 (1946).
44. Chwang, A. T. & Wu, T. Y.-T. Hydromechanics of low-Reynolds-number flow. Part 2. Singularity method for Stokes flows. *J. Fluid Mech.* **67**, 787–815 (1975).
45. Viola, I. M., Bot, P. & Riotte, M. On the uncertainty of CFD in sail aerodynamics. *Int. J. Numer. Methods Fluids* **72**, 1146–1164 (2013).



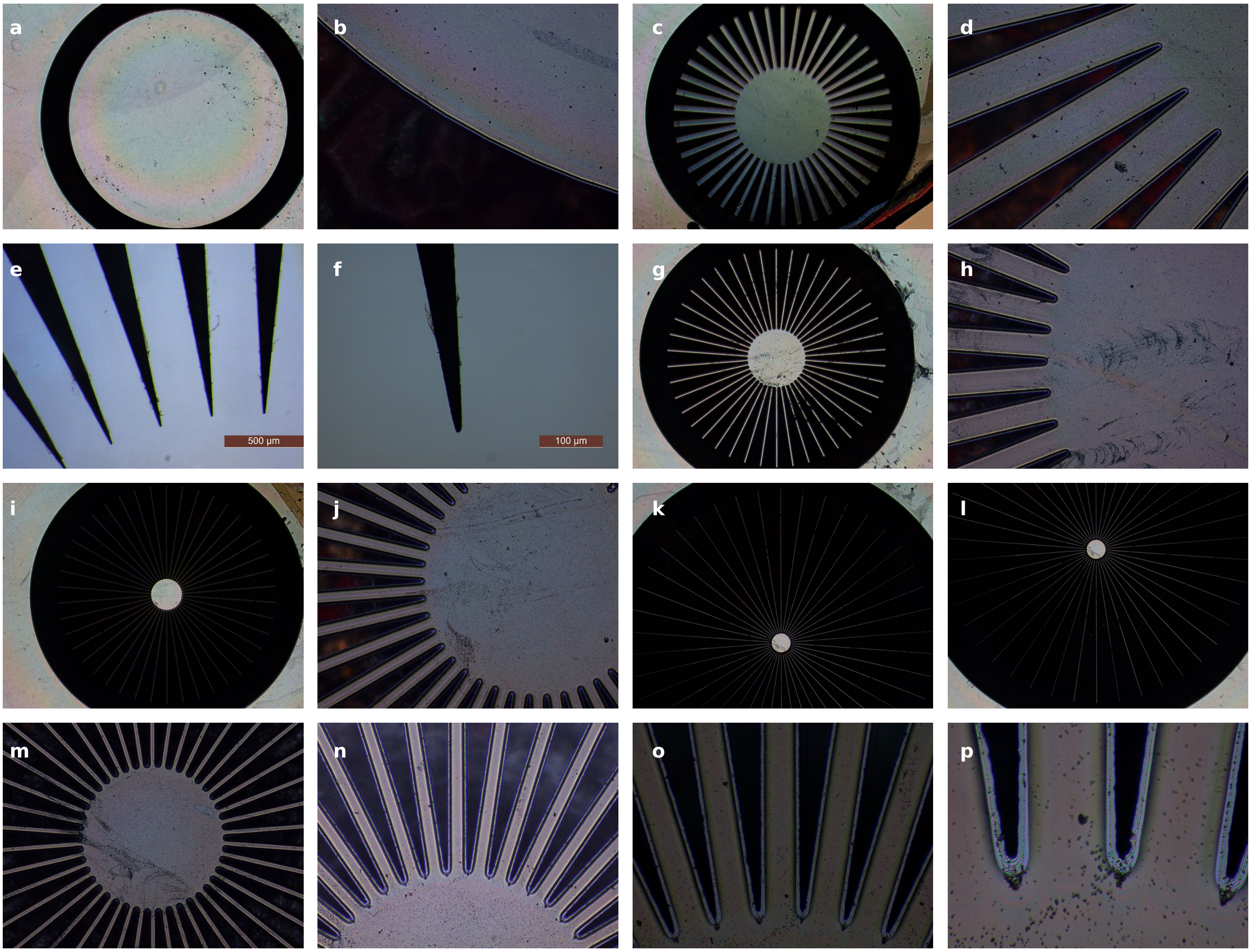
Extended Data Fig. 1 | SVR visualization of the wake of 10 fixed dandelion seeds. The flow speed is half of the terminal velocity of the seed. Each image was obtained using long-exposure photography.



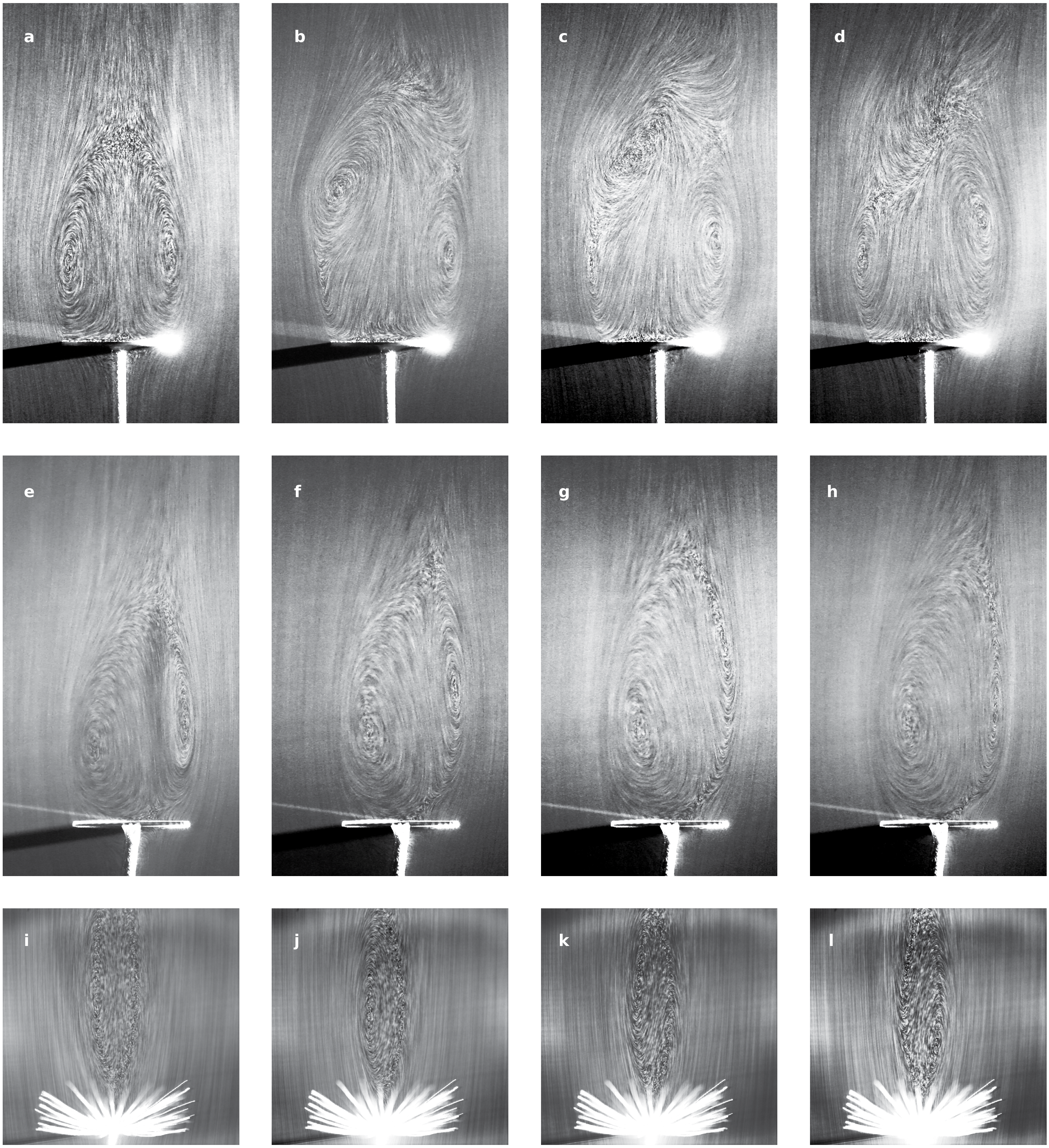
Extended Data Fig. 2 | SVR visualization of the wake of 10 freely flying dandelion seeds. a–j, Each image corresponds to a snapshot from a video of the flight of the dandelions in the wind tunnel. The images show the seeds as they pass through the laser sheet, and the SVR may be difficult to identify in some panels because of the orientation of the laser sheet with respect to the axis of the SVR.



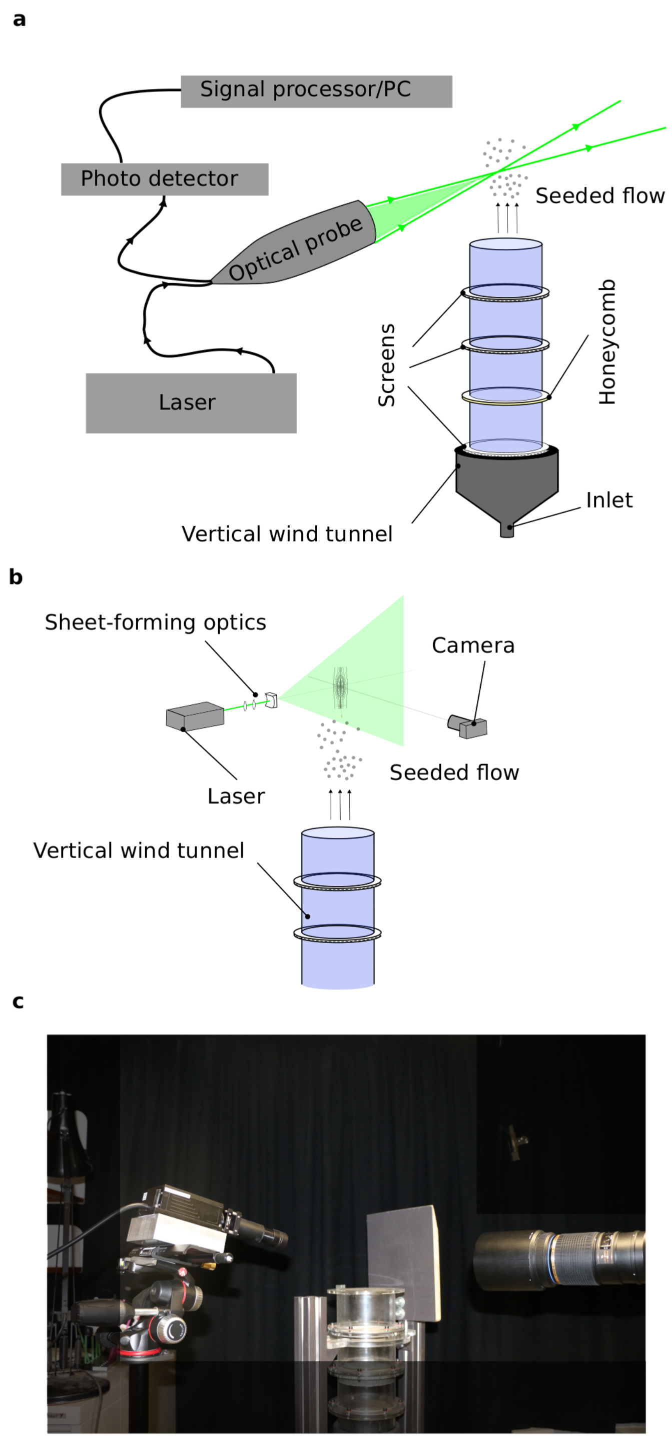
Extended Data Fig. 3 | The breakdown in symmetry in the SVR of dandelion seeds. a, **b**, At low speeds, the SVR is axisymmetric. **a**, Contrast-enhanced image. **b**, Original image. **c**, **d**, At higher speeds, this symmetry is lost. **c**, Contrast-enhanced image. **d**, Original image. **a**–**d**, Experiments were repeated independently on $n = 10$ biological samples, with similar results. **e**, **f**, The axisymmetry of SVR at low Re (**e**) breaks down at higher Re (**f**).



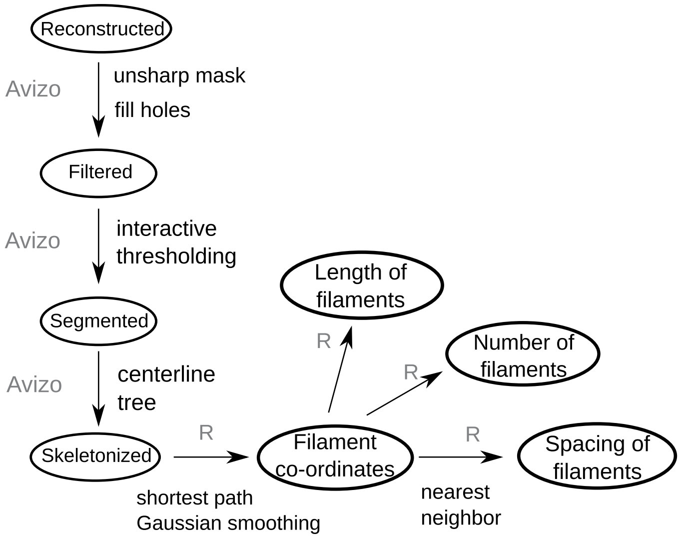
Extended Data Fig. 4 | Images of porous disks showing the resolution of the technique for disks of various porosities. a, **b**, Impervious disk. **c**–**f**, A disk with 33% porosity. **g**, **h**, A disk with 55% porosity. **i**, **j**, A disk with 75% porosity. **k**–**p**, A disk with 89% porosity.



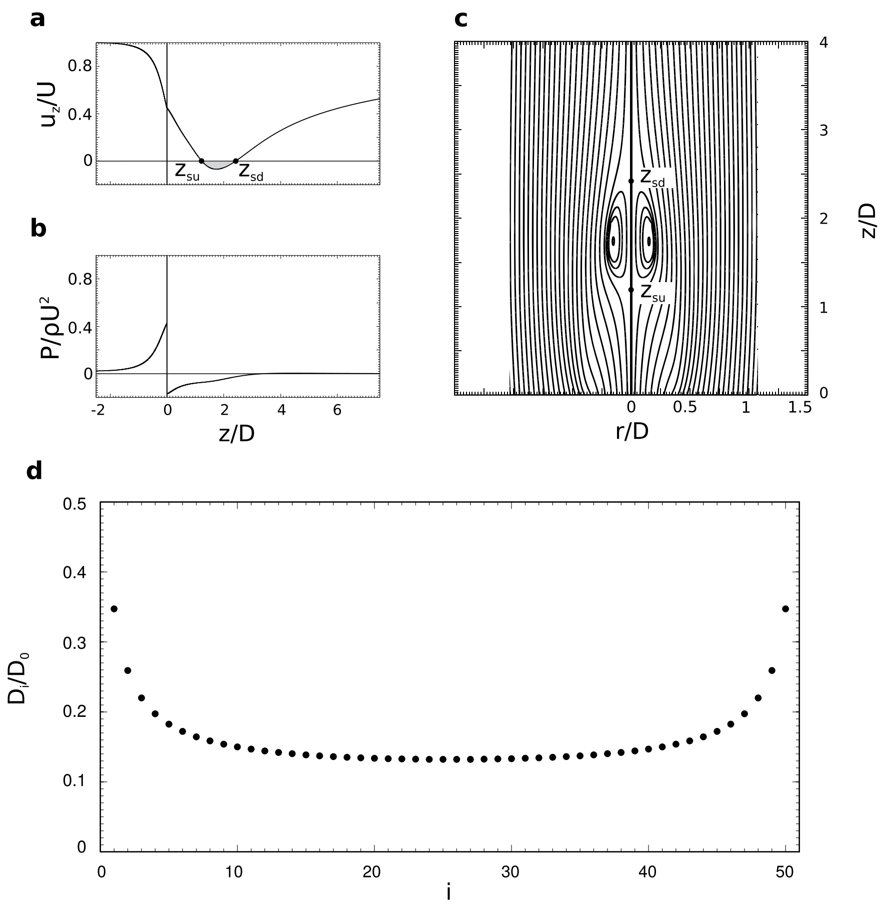
Extended Data Fig. 5 | Steady and unsteady wake behind porous disks and pappi. Video snapshots are shown. **a–d**, The flow visualization behind a solid disk, with a steady wake (**a**) and an unsteady wake at three time points within one period of vortex shedding (**b–d**). **e–h**, The flow around a porous disk ($\varepsilon = 0.75$) with a steady wake (**e**) and an unsteady wake at three time points within one period of vortex shedding (**f–h**). **i–l**, The wake behind a dandelion sample with a steady SVR (**i**) and at three time points within one period of vortex shedding (**j–l**).



Extended Data Fig. 6 | The experimental setup for laser Doppler anemometry and flow visualization. a, **b**, Schematic drawings of the experimental setup for laser Doppler anemometry to measure the flow speed and turbulent intensity in the wind tunnel (**a**) and the experimental setup for flow visualization in the wind tunnel using a CW laser and high-speed camera (**b**). **c**, Photograph of the actual experimental setup for flow visualization.



Extended Data Fig. 7 | Workflow for post-processing of the μCT scan data. Image processing workflow for analysis of μCT data indicating the algorithms performed and the software used (Avizo or R).



Extended Data Fig. 8 | The flow past a porous disk using direct numerical simulations and boundary integral methods. a–c, The axial velocity u_z/U (**a**), pressure $p/\rho U^2$ (**b**) and streamlines (**c**), showing the presence of an SVR with upstream and downstream stagnation points z_{su} and z_{sd}, respectively. **d**, The reduction in the drag force on filaments within an array moving at slow speeds calculated using a boundary integral method. The force D_i on the ith filament of a rectangular pappus, divided by the drag force for an isolated filament D_0.

Extended Data Table 1 | μCT scan-acquisition settings

X-ray energy	25 keV
X-ray power	14 W
Distance (X-ray to sample)	71 mm
Acquisition mode	reflectance
Camera type	Perkin-Elmer
Distance (camera to X-ray)	549.5 mm
Filter	none
Pixel size	0.2 mm
No. projections	2000
Exposure	2s

Extended Data Table 2 | Morphological data of dandelion seeds

	Diameter d (μm)	Length L (mm)	Filaments n	Porosity ϵ
mean =	16.3	7.41	100	0.916
CI =	15.7 − 17.0	7.35 − 7.46	95 − 106	0.907 − 0.923
n =	10	937	10	10

Data are shown as mean and 95% confidence intervals and the number of samples is shown.

https://doi.org/10.1038/s41586-018-0595-z

Cervical excitatory neurons sustain breathing after spinal cord injury

Kajana Satkunendrarajah[1,5]*, Spyridon K. Karadimas[2,5], Alex M. Laliberte[1], Gaspard Montandon[3,4] & Michael G. Fehlings[1,2]*

Dysfunctional breathing is the main cause of morbidity and mortality after traumatic injury of the cervical spinal cord[1,2] and often necessitates assisted ventilation, thus stressing the need to develop strategies to restore breathing. Cervical interneurons that form synapses on phrenic motor neurons, which control the main inspiratory muscle, can modulate phrenic motor output and diaphragmatic function[3–5]. Here, using a combination of pharmacogenetics and respiratory physiology assays in different models of spinal cord injury, we show that mid-cervical excitatory interneurons are essential for the maintenance of breathing in mice with non-traumatic cervical spinal cord injury, and are also crucial for promoting respiratory recovery after traumatic spinal cord injury. Although these interneurons are not necessary for breathing under normal conditions, their stimulation in non-injured animals enhances inspiratory amplitude. Immediately after spinal cord injury, pharmacogenetic stimulation of cervical excitatory interneurons restores respiratory motor function. Overall, our results demonstrate a strategy to restore breathing after central nervous system trauma by targeting a neuronal subpopulation.

The cervical spinal cord contains the white matter tracts, pre-phrenic interneurons, and phrenic motor neurons (PMNs) that control the diaphragm. High cervical spinal cord injury (SCI) denervates the descending excitatory drive from rostral ventral respiratory group neurons to PMNs, leading to notable respiratory dysfunction[6,7]. Respiratory insufficiency and subsequent mechanical ventilation in the first few days after SCI often lead to complications such as atelectasis, pneumonia and ventilator-dependent respiratory failure, and 80% of deaths among patients with SCI are the result of respiratory complications[1,2,8,9]. High cervical hemisection has been the predominant model used to investigate the respiratory consequences of SCI[10]. Similar to human patients with SCI, rodents receiving a C2 hemisection (C2Hx) demonstrate acute paralysis of the ipsilateral hemidiaphragm with decreased ventilation.

Non-traumatic cervical spinal cord injury (ntSCI) encompasses a variety of spinal diseases such as spinal degeneration, infection and tumour or vascular disorders that cause injury to the spinal cord without major trauma[11]. The most frequent underlying condition is cervical myelopathy[12], which is associated with progressive, chronic compression of the cervical spinal cord[13] due to spinal degeneration. In contrast to SCI, ntSCI in cervical myelopathy does not result in notable respiratory deficits[14,15] despite substantial cervical spinal cord damage, loss of motor neurons[16–18] and severe neurological deficits[19,20]. Understanding the changes in cervical microcircuitry that enable the preservation of breathing in ntSCI may uncover potential targets for the restoration of breathing after SCI. To this end, ntSCI was modelled in its commonest form, cervical myelopathy.

As with humans (Extended Data Fig. 1), ntSCI mice with severe progressive cervical cord compression (Extended Data Fig. 2) have marked impairment in gait and forelimb function (Extended Data Figs. 3, 4). However, ntSCI mice had similar arterial blood gases to sham mice (Fig. 1a), indicating adequate ventilation. Akin to humans with ntSCI of the cervical cord[14], respiratory assessment during quiet normoxic breathing demonstrated only mild respiratory changes, a 28.6% reduction in inspiratory duration, without significant changes in inspiratory amplitude or area (Fig. 1b–d). These mild changes in respiratory function occurred despite the loss of 64% of cholera toxin b (CTb)-traced PMNs (Fig. 1e). However, the surviving PMNs received increased excitatory input, indicated by the increased number of vesicular glutamate transporter 2 (VGLUT2)$^+$ boutons on CTb$^+$ PMNs, as spinal cord compression progressively increased over time (Fig. 1f). Eight weeks after the induction of ntSCI, when the loss of PMNs was most notable, the number of VGLUT2$^+$ boutons on PMNs was 5.4 times greater than the sham values.

Notably, ntSCI mice maintained robust inspiratory-related diaphragmatic electromyography (EMG) activity on the ipsilateral hemidiaphragm immediately after a C2Hx that disrupted all ipsilateral excitatory drive to the associated PMN pool (Extended Data Figs. 5, 6). This activity was rhythmic and synchronous with the contralateral side (Extended Data Fig. 5a, b). By contrast, sham mice had an absence of inspiratory-related EMG activity on the ipsilateral hemidiaphragm immediately after C2Hx. Trans-synaptic tracing from the hemidiaphragm of ntSCI and sham mice showed that the number of cervical interneurons connecting to the preserved PMNs in the ntSCI mice increased 2.75-fold compared to sham (Extended Data Fig. 7a–d). These interneurons resided mainly in the intermediate grey matter contralateral to the injection side at the C4–C5 level. Together, these findings suggest that the recruitment of an excitatory neural network at the midcervical level preserves breathing in ntSCI.

To show that the preservation of respiratory function in ntSCI was the result of the recruitment of cervical excitatory interneurons (eINs), we selectively expressed the chimaeric receptor/chloride channel (PSAM^{L141F}–GlyR; hereafter referred to as PSAM) in these cells (AAV-FLEX-PSAM-GlyR.GFP viral vector injected in *Vglut2::cre-tdTomato* mice; *Vglut2* is also known as *Slc17a6*) (Fig. 2a, b). PSAM reduces neuronal activity in response to its ligand PSEM^{308}. VGLUT2 is the main glutamate transporter in the spinal cord[21], and selective expression of Cre recombinase in only VGLUT2$^+$ interneurons combined with the spatial specificity of our injection resulted in PSAM expression only in the midcervical eINs and not in motor neurons (hereafter referred to as VGLUT2-PSAM mice; Fig. 2c and Extended Data Fig. 8). VGLUT2-PSAM mice then received ntSCI or sham surgery (VGLUT2-PSAM ntSCI or VGLUT2-PSAM sham mice, respectively; Fig. 2a).

Eight weeks after the induction of ntSCI, transient silencing of mid-cervical eINs with the artificial ligand PSEM^{308} decreased inspiratory amplitude by 33% (Fig. 2d, e). Silencing these neurons increased the average number of hypopnoea events (breaths with an inspiratory amplitude less than 50% of the control value), from 5.24% to 61.17% of the total recorded breaths (Fig. 2e). These events occasionally led to short periods of indiscernible low inspiratory activity. By contrast, the administration of PSEM^{308} in VGLUT2-PSAM sham mice did not affect respiratory function (Fig. 2e). Therefore, although mid-cervical eINs are unnecessary for normal breathing, they are crucial

[1]Krembil Research Institute, University Health Network, Toronto, Ontario, Canada. [2]Division of Neurosurgery, Department of Surgery, University of Toronto, Ontario, Canada. [3]St Michael's Hospital, Toronto, Ontario, Canada. [4]Department of Medicine, University of Toronto, Ontario, Canada. [5]These authors contributed equally: Kajana Satkunendrarajah, Spyridon K. Karadimas. *e-mail: kajanasatkune@gmail.com; michael.fehlings@uhn.ca

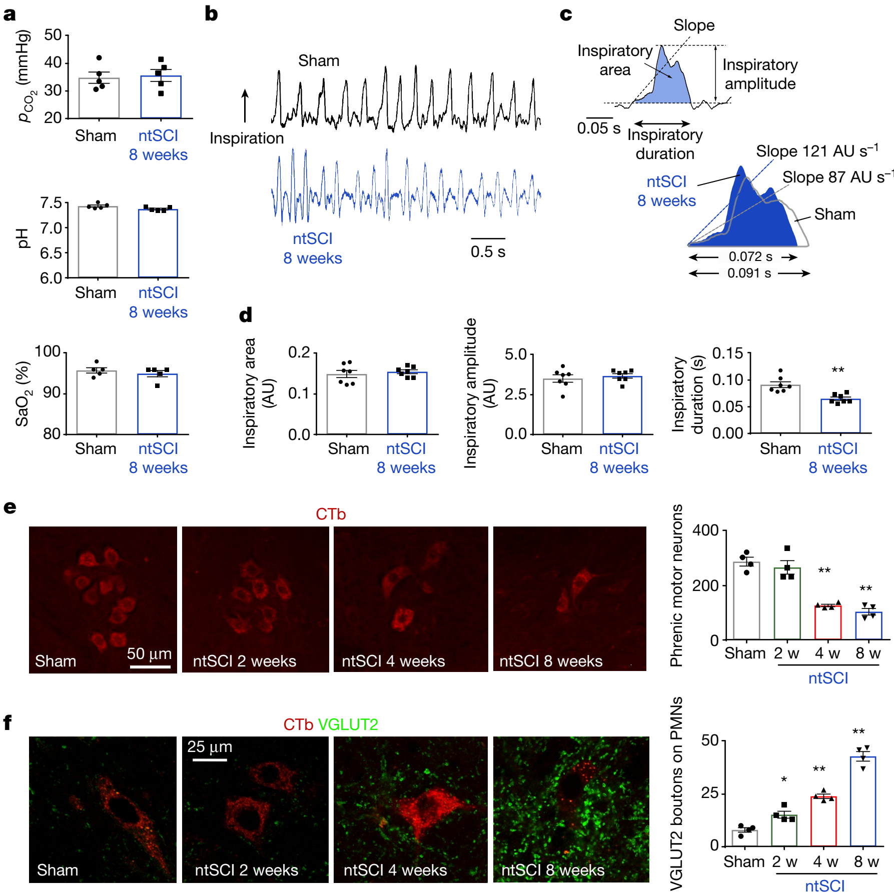


Fig. 1 | Adequate ventilation in ntSCI despite loss of PMNs. a, Sham and ntSCI mice had similar arterial CO_2 partial pressure (p_{CO2}; $P = 0.791$, 95% confidence interval (CI) = −6.082 to 7.722), oxygen saturation (SaO_2; $P = 0.455$, 95% CI = −3.152 to 1.552), and pH ($P = 0.065$; 95% CI = −0.1240 to 0.0048). P values determined by two-sided t-test; $n = 5$ mice per group. **b**, Representative plethysmography tracings from sham and ntSCI mice. **c**, Annotated diagrams demonstrating plethysmography waveforms corresponding to single inhalations from sham and ntSCI mice. **d**, ntSCI mice had reduced inspiratory duration (**$P = 0.0015$, 95% CI = −0.038 to −0.012), but lacked significant differences in inspiratory area ($P = 0.5403$, 95% CI = −0.016 to 0.029) or amplitude ($P = 0.5615$, 95% CI = −0.431 to 0.757). P values determined by two-sided t-test; $n = 7$ mice per group. AU, arbitrary units. **e**, After 4 and 8 weeks, ntSCI mice showed decreased numbers of PMNs (CTb^+) compared to sham mice ($P = 2.54 \times 10^{-4}$, 95% CI = 93.53 to 229.5 (4 weeks), and $P = 6.86 \times 10^{-5}$, 95% CI = 116.5 to 252.5 (8 weeks)). P values determined by one-way ANOVA with Tukey's post hoc test; $n = 4$ mice per group. **f**, After ntSCI, there was a progressive increase in the number of $VGLUT2^+$ presynaptic boutons (green) on preserved PMNs (CTb^+, red) (2 weeks: $P = 0.033$, 95% CI = −13.96 to −0.5357; 4 weeks: $P = 2.50 \times 10^{-4}$, 95% CI = −22.71 to −9.286; 8 weeks: $P = 5.56 \times 10^{-8}$, 95% CI = −41.71 to −28.29). P values determined by one-way ANOVA with Tukey's post hoc test; $n = 4$ mice per group. Data are mean ± s.e.m. *$P < 0.05$; **$P < 0.01$.

for maintaining adequate ventilation in ntSCI. To examine whether mid-cervical eINs can maintain diaphragmatic function in the absence of ipsilateral bulbospinal input, a C2Hx was performed (Extended Data Fig. 6) on VGLUT2-PSAM ntSCI mice. VGLUT2-PSAM ntSCI animals had rhythmic respiratory-related activity in the ipsilateral hemidiaphragm, synchronous with the contralateral hemidiaphragm activity, despite ipsilateral C2Hx (Fig. 2f). Acute administration of $PSEM^{308}$ abolished this respiratory-related EMG activity recorded on the ipsilateral diaphragm. Therefore, mid-cervical eINs are required to sustain ipsilateral diaphragm activity in the absence of direct ipsilateral bulbospinal drive, probably functioning as relay neurons providing descending inspiratory drive.

Partial spontaneous recovery of respiratory function occurs 2–4 weeks after traumatic SCI (C2Hx)[22–24]. This recognized form of respiratory plasticity is thought to occur via activation of latent pathways[25,26]. The fact that cervical eINs preserve breathing in ntSCI makes them strong candidates to mediate spontaneous respiratory plasticity in SCI. To examine this hypothesis, SCI was performed in VGLUT2-PSAM mice (Fig. 3a, b). Acute silencing of eINs with $PSEM^{308}$ four weeks after SCI resulted in substantial disruption of the animals' ability to maintain breathing (Fig. 3c), indicating their crucial role in respiratory plasticity after SCI (Fig. 3c). When SCI mice were able to generate inspiratory activity, the mean amplitude was 69.5% lower than baseline levels (Fig. 3d). Although the cervical eINs are unnecessary for normal breathing, they are recruited to sustain breathing after PMN loss and/or lack of excitatory drive from the ventral medulla to the spinal respiratory circuitry.

To assess the functional connectivity between the cervical eINs and the spinal respiratory network in uninjured mice during spontaneous breathing, we selectively stimulated this neuronal population using a chemogenetic approach based on designer receptors exclusively activated by designer drugs (DREADDs). The adeno-associated virus AAV5-DIO-hM3Dq-mCherry was injected into the cervical cord of *Vglut2::cre* mice (hereafter referred to as VGLUT2-hM3Dq; Extended Data Fig. 9a, b). Administration of the hM3Dq-specific agonist clozapine *N*-oxide (CNO) to these mice resulted in a 28.5% increase in inspiratory amplitude compared to saline, without changing inspiratory frequency (Extended Data Fig. 9c, e). The absence of changes to respiratory frequency indicated that the stimulation of cervical eINs, in contrast to epidural stimulation of the mid-cervical spinal cord[27], does not affect the supraspinal centres that modulate respiratory pattern. Moreover, saline administration did not change either the inspiratory amplitude or frequency—excluding the possibility that the intraperitoneal injection, per se, is responsible for the changes observed after

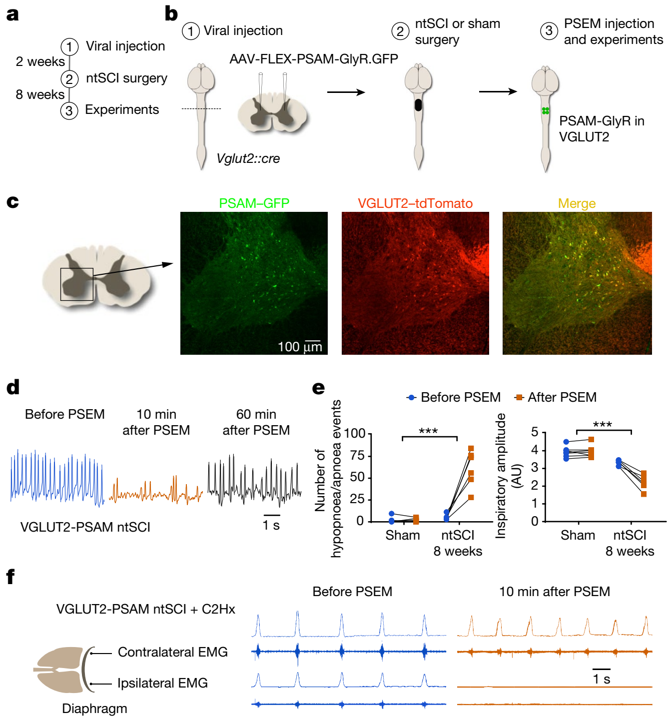


Fig. 2 | Cervical eINs are necessary for maintaining breathing after ntSCI. a, **b**, PSAM^{L141F}–GlyR–GFP was expressed in eINs of the cervical intermediate grey matter of VGLUT2-PSAM mice. Mice underwent ntSCI or sham surgery. **c**, Representative confocal images demonstrating the expression of PSAM (GFP^{+}) in only eINs (VGLUT2–tdTomato^{+}) of the intermediate grey matter area (6 biological replicates). **d**, Representative plethysmographic tracings from VGLUT2-PSAM ntSCI mice before, and 10 and 60 min after intraperitoneal injection of the ligand PSEM^{308}. **e**, PSEM^{308} significantly increased the number of hypopnoea events in VGLUT2-PSAM ntSCI animals, whereas there was no change in VGLUT2-PSAM sham mice. The interaction between injury and the change in hypopnoea events after PSEM^{308} injection was highly significant ($P = 2.78 \times 10^{-5}$, repeated-measures ANOVA). Similar to the hypopnoea events, the interaction between injury and change in inspiratory amplitude after PSEM^{308} injection was highly significant ($P = 1.11 \times 10^{-5}$, repeated-measures ANOVA), supporting the marked decrease in inspiratory amplitude after PSEM^{308} injection in VGLUT2-PSAM ntSCI mice ($n = 7$ mice for sham; $n = 6$ mice for ntSCI). **f**, VGLUT2-PSAM ntSCI mice had respiratory-related diaphragmatic EMG activity on the ipsilateral hemidiaphragm immediately after a C2Hx that was rhythmic and synchronous with the contralateral side. Inspiratory burst activity was completely abolished 10 min after PSEM^{308} administration, whereas the right hemidiaphragmatic activity persisted in all PSAM ntSCI mice. Top traces, integrated EMG; bottom trace, raw EMG tracings.

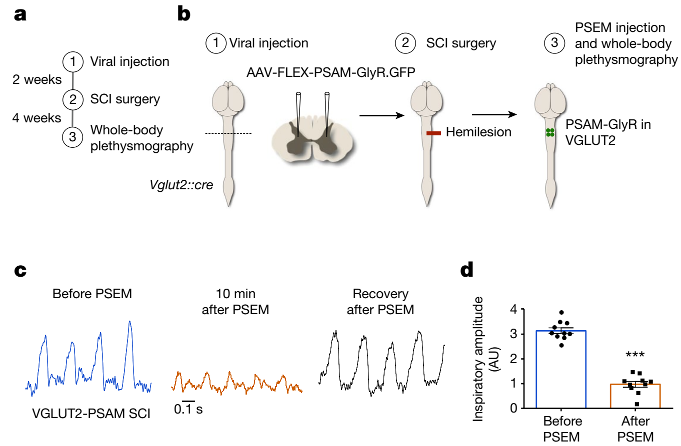


Fig. 3 | Cervical glutamatergic cells are crucial for sustaining breathing after traumatic cervical SCI. a, **b**, Experimental scheme used to express PSAM in eINs of cervical intermediate grey matter in SCI. **c**, Representative plethysmographic tracings from VGLUT2-PSAM SCI mice before, and 10 and 60 min after intraperitoneal administration of PSEM^{308}, demonstrating the animals' inability to breathe after silencing the cervical eINs. **d**, The inspiratory amplitude after PSEM^{308} injection was significantly reduced from baseline in the 4-week VGLUT2-PSAM SCI mice. $^{***}P = 1.06 \times 10^{-8}$, 95% CI $= -2.418$ to -1.918; paired two-sided *t*-test; $n = 10$ mice. Data are mean ± s.e.m.

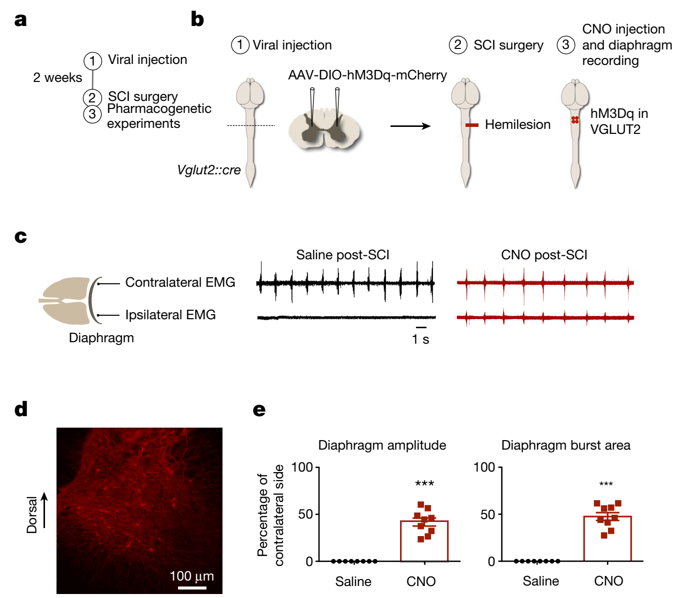


Fig. 4 | Stimulation of cervical eINs restores respiratory function immediately after SCI. a, **b**, Experimental scheme used to stimulate the cervical eINs of the C4–C5 level immediately after SCI. **c**, Representative EMG recordings for VGLUT2-hM3Dq SCI animals after either saline or CNO injection. **d**, Representative confocal image demonstrating the expression of hM3Dq–mCherry in glutamatergic cells of the area of interest (9 biological replicates). No expression was identified in anterior horn regions containing the PMNs. **e**, Bar graphs demonstrating peak amplitude and area of ipsilateral diaphragmatic inspiratory bursting, expressed as a percentage of the contralateral (intact) side, immediately after traumatic SCI in both saline- and CNO-treated groups. All VGLUT2-hM3Dq SCI mice receiving CNO showed rhythmic diaphragmatic EMG activity ipsilateral to the injury immediately after SCI (41.78% and 47.5% of the contralateral EMG amplitude and area, respectively); no control VGLUT2-hM3Dq mice receiving saline showed any activity on the injured side. $^{***}P = 8.23 \times 10^{-5}$ for both amplitude and area, two-sided Mann–Whitney-*U* test; $n = 8$ mice for SCI plus saline group, and $n = 9$ mice for SCI plus CNO group. Data are mean ± s.e.m.

administration of CNO (Extended Data Fig. 9c, e). Finally, no changes in inspiratory amplitude were found after CNO administration in naive mice (Extended Data Fig. 9d, f). Hence, the stimulation of mid-cervical eINs enhances the respiratory motor output under unanaesthetized normal breathing conditions without modulating supraspinal centres.

As previously noted, respiratory insufficiency in the acute phase of traumatic cervical SCI often necessitates mechanical ventilation, and can ultimately lead to life-threatening respiratory complications[1,2]. Effective treatments to restore breathing during this crucial phase are currently lacking. Because the activation of mid-cervical eINs enhanced breathing in the uninjured state, this population represents an ideal target to rescue breathing at the acute phase after SCI. Therefore, we performed a left C2Hx on VGLUT2-hM3Dq mice (Fig. 4a, b, d) and promptly stimulated the mid-cervical eINs, resulting in immediate motor recovery of the left hemidiaphragm to 41.78% and 47.5% of the contralateral EMG amplitude and area, respectively (Fig. 4e). Despite the complete disruption of the ipsilateral bulbospinal connections, this activity was rhythmic and synchronous with the contralateral

(uninjured) side (Fig. 4c). By contrast, when the SCI mice received saline, they did not display any respiratory-related hemidiaphragmatic activity on the side of the C2Hx (Fig. 4c, e).

In conclusion, we show that cervical eINs have a crucial role in facilitating respiratory function under injury states that directly damage the PMNs or disrupt descending inputs to PMNs. Furthermore, stimulation of cervical eINs after traumatic SCI is sufficient to rescue breathing in mice at the acute stage, the most critical period for patients with SCI. Therefore, viral vectors can be used to target this subpopulation in humans with CNS disorders that affect respiratory function such as amyotrophic lateral sclerosis and SCI. In addition, our findings provide insights into the role of cervical eINs in breathing. Although not essential to normal breathing, the increase in inspiratory function produced by their stimulation suggests that these interneurons connect to PMNs and are recruited to modulate inspiratory output. Although these neurons can influence PMN function under specific physiological conditions[28], the modest and gradual pace of respiratory recovery after SCI[25,26] suggest a limited capacity for rapid modulation of PMN function after trauma. As such, this study opens new possibilities for research in motor control of respiration, providing a compelling anatomical substrate for acute respiratory intervention and questions about the function of these neurons in the uninjured state.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0595-z.

Received: 26 July 2017; Accepted: 14 August 2018;
Published online 10 October 2018.

1. Jackson, A. B. & Groomes, T. E. Incidence of respiratory complications following spinal cord injury. *Arch. Phys. Med. Rehabil*. **75**, 270–275 (1994).
2. Berlly, M. & Shem, K. Respiratory management during the first five days after spinal cord injury. *J. Spinal Cord Med*. **30**, 309–318 (2007).
3. Palisses, R., Perségol, L. & Viala, D. Evidence for respiratory interneurones in the C3-C5 cervical spinal cord in the decorticate rabbit. *Exp. Brain Res*. **78**, 624–632 (1989).
4. Lane, M. A. et al. Cervical prephrenic interneurons in the normal and lesioned spinal cord of the adult rat. *J. Comp. Neurol*. **511**, 692–709 (2008).
5. Cregg, J. M. et al. A latent propriospinal network can restore diaphragm function after high cervical spinal cord injury. *Cell Reports* **21**, 654–665 (2017).
6. Como, J. J. et al. Characterizing the need for mechanical ventilation following cervical spinal cord injury with neurologic deficit. *J. Trauma* **59**, 912–916, discussion 916 (2005).
7. Lane, M. A. et al. Respiratory function following bilateral mid-cervical contusion injury in the adult rat. *Exp. Neurol*. **235**, 197–210 (2012).
8. Brown, R., DiMarco, A. F., Hoit, J. D. & Garshick, E. Respiratory dysfunction and management in spinal cord injury. *Respir. Care* **51**, 853–868, discussion 869–870 (2006).
9. DeVivo, M. J., Black, K. J. & Stover, S. L. Causes of death during the first 12 years after spinal cord injury. *Arch. Phys. Med. Rehabil*. **74**, 248–254 (1993).
10. Alilain, W. J., Horn, K. P., Hu, H., Dick, T. E. & Silver, J. Functional regeneration of respiratory pathways after spinal cord injury. *Nature* **475**, 196–200 (2011).
11. Grassner, L. et al. Nontraumatic spinal cord injury at the neurological intensive care unit: spectrum, causes of admission and predictors of mortality. *Ther. Adv. Neurol. Disord*. **9**, 85–94 (2016).
12. New, P. W. et al. International retrospective comparison of inpatient rehabilitation for patients with spinal cord dysfunction epidemiology and clinical outcomes. *Arch. Phys. Med. Rehabil*. **96**, 1080–1087 (2015).
13. Nouri, A., Tetreault, L., Singh, A., Karadimas, S. K. & Fehlings, M. G. Degenerative cervical myelopathy: epidemiology, genetics, and pathogenesis. *Spine* **40**, E675–E693 (2015).
14. Toyoda, H., Nakamura, H., Konishi, S., Terai, H. & Takaoka, K. Does chronic cervical myelopathy affect respiratory function? *J. Neurosurg. Spine* **1**, 175–178 (2004).
15. Bhagavatula, I. D. et al. Subclinical respiratory dysfunction in chronic cervical cord compression: a pulmonary function test correlation. *Neurosurg. Focus* **40**, E3 (2016).
16. Karadimas, S. K. et al. Riluzole blocks perioperative ischemia-reperfusion injury and enhances postdecompression outcomes in cervical spondylotic myelopathy. *Sci. Transl. Med*. **7**, 316ra194 (2015).
17. Karadimas, S. K. et al. A novel experimental model of cervical spondylotic myelopathy (CSM) to facilitate translational research. *Neurobiol. Dis*. **54**, 43–58 (2013).
18. Yu, W. R., Liu, T., Kiehl, T. R. & Fehlings, M. G. Human neuropathological and animal model evidence supporting a role for Fas-mediated apoptosis and inflammation in cervical spondylotic myelopathy. *Brain* **134**, 1277–1292 (2011).
19. Malone, A., Meldrum, D. & Bolger, C. Gait impairment in cervical spondylotic myelopathy: comparison with age- and gender-matched healthy controls. *Eur. Spine J*. **21**, 2456–2466 (2012).
20. Ono, K. et al. Myelopathy hand. New clinical signs of cervical cord damage. *J. Bone Joint Surg. Br*. **69**, 215–219 (1987).
21. Kullander, K. et al. Role of EphA4 and EphrinB3 in local neuronal circuits that control walking. *Science* **299**, 1889–1892 (2003).
22. Bezdudnaya, T., Hormigo, K. M., Marchenko, V. & Lane, M. A. Spontaneous respiratory plasticity following unilateral high cervical spinal cord injury in behaving rats. *Exp. Neurol*. **305**, 56–65 (2018).
23. Goshgarian, H. G. The crossed phrenic phenomenon and recovery of function following spinal cord injury. *Respir. Physiol. Neurobiol*. **169**, 85–93 (2009).
24. Alilain, W. J. & Goshgarian, H. G. MK-801 upregulates NR2A protein levels and induces functional recovery of the ipsilateral hemidiaphragm following acute C2 hemisection in adult rats. *J. Spinal Cord Med*. **30**, 346–354 (2007).
25. Golder, F. J. et al. Respiratory motor recovery after unilateral spinal cord injury: eliminating crossed phrenic activity decreases tidal volume and increases contralateral respiratory motor output. *J. Neurosci*. **23**, 2494–2501 (2003).
26. Fuller, D. D. et al. Modest spontaneous recovery of ventilation following chronic high cervical hemisection in rats. *Exp. Neurol*. **211**, 97–106 (2008).
27. Huang, R. et al. Modulation of respiratory output by cervical epidural stimulation in the anesthetized mouse. *J. Appl. Physiol*. **121**, 1272–1281 (2016).
28. Streeter, K. A. et al. Intermittent hypoxia enhances functional connectivity of midcervical spinal interneurons. *J. Neurosci*. **37**, 8349–8362 (2017).

Acknowledgements This research was supported by grant 2987 from the PVA Research Foundation (M.G.F. and K.S.), CIHR Grant MOP13683 (M.G.F.), AOSPINE Young Investigator Research Grant (K.S.); the Halbert Chair (M.G.F.) and the DeZwirek Foundation (M.G.F.). S.K.K. was supported by the Onassis Foundation. We thank A. Andreopoulou and J. Kallitsis for the ntSCI compression material; J. Austin for logistical support; C. Castro and L. Li for technical assistance; L. Teves (M. Tymianski's laboratory) for assistance with blood gas measurement. We thank our colleagues for constructive feedback.

Reviewer information *Nature* thanks J. Feldman and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Study conception: K.S., S.K.K. and M.G.F. Experimental design: K.S. and S.K.K. ntSCI surgeries: S.K.K. (A.M.L. for Extended Data Fig. 2). SCI surgeries: K.S. Viral intraspinal injections: K.S. and S.K.K. Electrophysiological experiments and analysis: K.S. Plethysmography: K.S. and G.M. Anatomical experiments: S.K.K. and K.S. Neurobehavioural and blood gas experiments: K.S. and S.K.K. Molecular and viral work: A.M.L. Statistics: A.M.L. Supervision: M.G.F. All authors discussed the results and participated in writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0595-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0595-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to K.S. or M.G.F.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mouse lines. All experiments were performed in accordance with Canadian Council on Animal Care guidelines and were approved by University Health Network's animal research committee. C57BL/6J mice (Jackson Labs 000664), *Vglut2::cre* mice (B6-*Slc17a6*$^{tm2(cre)Lowl}$/J, Jackson Labs 016963), and Ai9 mice (B6.129S-*Gt(ROSA)Sor26*$^{tm9(CAG\text{-}tdTomato)Hze}$/J Jackson Labs 007908) were purchased from commercial vendors. Female mice, aged 8–10 weeks, were used for all experiments.

Acute cervical spinal cord injury. A C2 hemisection injury was performed as previously described[29]. In brief, mice were anaesthetized with 1.5% isoflurane in oxygen (1 l min^{-1}). Once the mice reached a surgical plane of anaesthesia, under aseptic conditions the mice were placed in a stereotaxic frame (Kopf Instruments) on a heating pad maintained at 37 °C. Following a 1.5-cm dorsal midline skin incision, the paravertebral muscles were retracted to expose the dorsal aspect of the cervical spinal cord at the C2 vertebral level. Following a durotomy and visualization of the C2 dorsal roots, the cord was hemisected caudal to the C2 roots with a microblade. The microblade was placed at the midline and extended towards the lateral most border. The hemisection was considered to be complete by the contact of the microblade with the ventral surface of the spinal canal, lack of tissue sparing and a clear separation of the spinal cord stumps.

ntSCI surgery. In brief, a sheet of polyaromatic ether was surgically inserted under the C4–C6 laminae in isoflurane anaesthetized mice, ultimately resulting in progressive ossification and compression of the spinal cord[30]. Following the material insertion, the surgical incision site was sutured, and buprenorphine was administered peri-operatively for analgesia.

Blood gas analysis. Blood samples (0.3 ml) were drawn from the left ventricle of briefly anaesthetized (isoflurane) ntSCI and sham operated mice. The blood was collected in heparin-coated syringes and immediately transferred to a CG4+ cartridge (Abaxis) to measure oxygen and carbon dioxide partial pressure (p_{O2} and p_{CO2}), saturation of oxygen (SaO_2) and pH using the Vetscan I-Stat-1 system (Abaxis).

CNS viral injections. For all CNS injections, mice were anaesthetized using isofluorane. Injections were performed using 1 μl pulled glass pipets, connected to a microinjector (picospritzer III, Parker) and attached to a stereotaxic apparatus (Model 940 Small Animal Stereotaxic Instrument with digital display, Kopf Instruments). A total volume of 0.5–0.7 μl of virus was injected at any individual site at a flow rate of 100 nl s^{-1}. After injection, the pipette was kept in place for 10 min and slowly withdrawn to prevent backflow. The skin was sutured and buprenorphine (0.05 mg kg^{-1}) was administered peri-operatively for analgesia. For confirmation of injection efficiency, mice were transcardially perfused with 4% paraformaldehyde (PFA) in PBS and spinal cord tissue was dissected and post-fixed in 4% PFA and 30% sucrose solution for cryosectioning and subsequent immunohistochemistry at the conclusion of experiments.

Pharmacogenetic inhibition of cervical excitatory neurons. To transiently silence the midcervical excitatory neurons of the intermediate grey matter, we injected adeno-associated viruses carrying *cre*-dependent inhibitory PSAM (AAVDJ-syn::FLEX-rev::PSAM^{L141F}-GlyR-IRES-GFP, 1.0×10^{12} genome copies (GC) per ml, plasmid was a gift from S. Sternson; Addgene plasmid 32479) bilaterally in *Vglut2::Cre-tdtomato* mice (Ai9) at the C4 and C5 spinal levels using a picospritzer. The following coordinates were used: 0.5 mm lateral and 0.7 mm ventral from the dorsal surface. Mice were allowed 2 weeks of rest after injections to allow full expression of constructs. For silencing experiments, PSEM^{308} in saline was administered intraperitoneally (5 mg kg^{-1}) and behavioural effects were measured for 1 h after injection.

Pharmacogenetic stimulation of cervical excitatory neurons. We injected adeno-associated viruses carrying *cre*-dependent *hM3Dq* excitatory DREADD (AAV5-DIO-hM3Dq-mCherry, 3.8×10^{12} GC per ml, UNC Vector Core) bilaterally in *Vglut2::cre* mice at the C4 and C5 spinal levels using a picospritzer using the same coordinates described in the previous paragraph. To stimulate the midcervial eINs, CNO in saline was administered intraperitoneally (0.2 mg kg^{-1}) and behavioural effects were measured for 2 h after injection.

Retrograde labelling of PMNs. Cholera toxin b Alexa Fluor 594 conjugate (CTb-594) was used to retrogradely label the PMNs before inducing ntSCI or sham surgery. This technique has been used to successfully label the PMNs[31]. In brief, awake animals were lightly restrained and tilted. A 5-μl Hamilton syringe was used to inject 10 μl of 0.2% CTb-594 solution (Molecular Probes). Injections were made into the thoracic cavity bilaterally via the fifth intercostal space. After the injections, animals were monitored for any evidence of respiratory distress.

Viral retrograde labelling of PMNs and prephrenic interneurons. PMNs and prephrenic interneurons were retrogradely labelled using PRV152 (NIH Center for Neuroanatomy with Neurotropic Viruses, 1.5×10^9 p.f.u. ml^{-1}) expressing a GFP. Following a laparotomy, the diaphragm muscle was exposed and 10 μl of the tracer was injected into one half of the diaphragm at multiple sites. Animals were euthanized 64 h after tracer injection to label ipsilateral PMNs and prephrenic interneurons.

Representative human data. Collection of representative human ntSCI data was performed with the consent of participants during the course of ongoing clinical studies approved by the University Health Network's Institutional Research Ethics Board. Data collection was performed in accordance with all applicable laws and policies governing human research and data privacy.

Gait analysis. Representative gait tracings from sham and ntSCI mice were collected using the Catwalk XT system (Noldus). Human gait parameters were collected using the GAITRite system, a validated[32], electronic pressure-sensitive walkway capable of recording footfalls and calculating spatiotemporal gait parameters. Representative gait traces were generated using the GAITRite Software based on a single ntSCI patient with moderate to severe deficits and an age-matched healthy control subject. Both individuals were instructed to walk across the GAITRite at a comfortable self-selected pace.

Horizontal ladder walk. For the horizontal ladder walk, two plexiglass plates (69 × 9 cm) were connected by 10 rungs, each 1 cm apart at the beginning and end of the horizontal ladder followed by rungs that were 2 cm apart in the middle[33]. The camera was set to 100 frames per second with a resolution of 1,280 × 1,024. Custom-made reflective markers were attached to the right shoulder, elbow and wrist. Forelimb movements and manual dexterity were tracked during targeted reaching and grasping behaviour while crossing the horizontal ladder using MaxTRAQ 2D software and analysed using MaxMate motion analysis toolbox for Excel (Innovations Systems).

Pellet reaching task. The pellet reaching task was performed as previously described[34]. In brief, mice were food-restricted to maintain 90% of their body weight. The task was performed in a chamber constructed from clear Plexiglas (1-mm thickness; dimensions 20 × 8.5 × 15 cm). The chamber also contained a single vertical slit that was 0.5-cm wide and 13-cm high through which the animals were trained to reach for a cheese pellet. The pellets were placed outside the slit on a 1.5-cm high platform. After three days of training, the animals were allowed to reach for pellets for 15 min. The reaching and grasping events were classified as either misses, when the animal makes no contact with the pellet during a reaching attempt, or successes. Success was defined when the mouse was able to successfully grasp the pellet and return it to the chamber.

Capellini handling test. The Capellini handling test was used to characterize forepaw dexterity while eating a piece of pasta. The animals were placed on a mirror in a clear box and uncooked pasta pieces (diameter 0.9 mm and length 26 mm) were eaten in three sequential trials after training. The time required to eat and the number of forepaw adjustments made while eating the pasta were quantified using a camera set to 100 frames per second with a resolution of 1,280 × 1,024.

Diaphragmatic EMGs recordings. Respiratory motor function was assessed using diaphragmatic EMGs as previously described[29,35]. In brief, mice were anaesthetized using isoflurane, and animal temperature was carefully maintained using a heating pad. An abdominal incision at the base of the rib cage was used to expose the caudal surface of the diaphragm muscle. Bipolar silver electrodes were inserted bilaterally into the diaphragm to record respiratory-related diaphragmatic EMGs. The acquired signal was differentially amplified at 10,000× and filtered (0.1–3 kHz) using a Cambridge Electronic Data acquisition system (CED1401) and Spike 2 software[10,29]. Raw diaphragmatic EMG activity was rectified and integrated using Spike2 software (Cambridge Electronic Design). Peak amplitude and burst area of the inspiratory bursts were determined via Spike2 for a period of 30 s in each animal. Peak values of left hemidiaphragmatic activity were standardized to the contralateral hemidiaphragmatic activity to determine the extent of recovery.

Whole-body plethysmography. Whole-body plethysmography was performed as previously described[36]. In brief, a differential pressure transducer (Data Sciences International) was calibrated before plethysmography measurements using a known airflow. Deflections of pressure caused by breathing were detected by the pressure transducer and preamplified using a Strain Gage Preamplifier Array and differentially amplified at 10,000× and filtered (0.1–3 kHz) using a Cambridge Electronic Data acquisition system (CED1401) and Spike 2 software. A bias airflow controller (Buxco) was used to maintain a steady flow of room air. Animals were provided sufficient time (1.5 h) to acclimatize to the plethysmography chamber (Buxco). After acclimatization, the mice became adapted to the environment and the rate of sniffing and activity was reduced allowing for experimentation. Mice were monitored for bouts of calm breathing over a period of 2 h, and periods when mice were sleeping, ambulating, grooming, or otherwise active were excluded from plethysmography analysis. Only periods of quiet resting were used for analysis. Single animals were injected with either saline, CNO, or PSEM^{308} and placed back into the plethysmography chamber. Periods of hypopnoea were quantified as previously described[37]; in brief, as periods of respiratory disturbance in which inspiratory amplitude was less than 50% of control values.

Immunohistochemistry. Cervical spinal cords were fixed for 2 h in 4% PFA in PBS, rinsed in PBS, cryoprotected in 30% (w/v) sucrose in PBS overnight and embedded in OCT mounting medium. Thirty-micrometre transverse sections were obtained on a cryostat. Sections were incubated overnight at 4 °C with the

primary antibody for vesicular glutamate transporter (VGLUT2, guinea pig polyclonal, 1:2,500, MilliporeSigma). Alexa Fluor-488-conjugated secondary antibody was incubated for 1 h at room temperature to visualize immunoreactive sites. Slides were rinsed, mounted in Vectashield mounting medium (Vectorlabs) and scanned on a LSM510 confocal microscope (Zeiss Microsystems). Multiple channels were scanned sequentially to limit fluorescence bleedthrough and false-positive signals.

Statistics and reproducibility. Sample size was selected based on previous experiments from our laboratory and others using the neurobehavioural and electrophysiological techniques described. In cases in which animals were allocated to distinct experimental groups, group identity was assigned at random. The experimenters were blinded to treatment (saline, CNO or PSEM[308]) for whole-body plethysmography, and to sample identity in the stereological cell counting of PMNs in ntSCI and sham mice. All animals that met experimental conditions were analysed, and no exclusion was required. All experiments were reliably reproduced, and the exact number (n) of biological replicates/mice is indicated in the figure legends. Results presented as representative experiments were repeated with similar results (specifically, Fig. 2c: 6 biological replicates; Fig. 4d: 9 biological replicates; Extended Data Fig. 2b: 4 biological replicates; Extended Data Fig. 6: 22 biological replicates; Extended Data Fig. 7d, e: 3 sham and 5 ntSCI mice; Extended Data Fig. 8: 6 biological replicates; Extended Data Fig. 9b: 8 biological replicates). Experiments presented in Extended Data Figs. 2a and 3 were not repeated for the purposes of this study, but represent typical results from our laboratory. All statistical tests were performed using SPSS v.21 (IBM), and graphs were generated using PRISM 6 (GraphPad Software). Graphs represent the mean $\pm$ s.e.m. unless noted otherwise. Assumptions of parametric statistical tests, such as normal data distribution and homoscedasticity (Brown–Forsythe test) were assessed in cases in which paired t-test, independent t-tests, one-way ANOVA and repeated-measures ANOVA were used. All statistical tests were two-tailed. Distributions of data were compared using independent samples t-test (Fig. 1a, d, Extended Data Figs. 4b, d–f and 7b), one-way ANOVA with Tukey's post hoc test (Fig. 1e, f), repeated-measures ANOVA with Sidak's post hoc test (Extended Data Fig. 7c), repeated-measures ANOVA (Fig. 2e and Extended Data Fig. 9c), paired t-test (Fig. 3d and Extended Data Fig. 9d), Mann–Whitney U-test (Fig. 4e), and the chi-square test for trend (Extended Data Fig. 5a). Test statistics, degrees of freedom, sample sizes and effect sizes for statistical analyses are reported in the Extended Data Table 1.

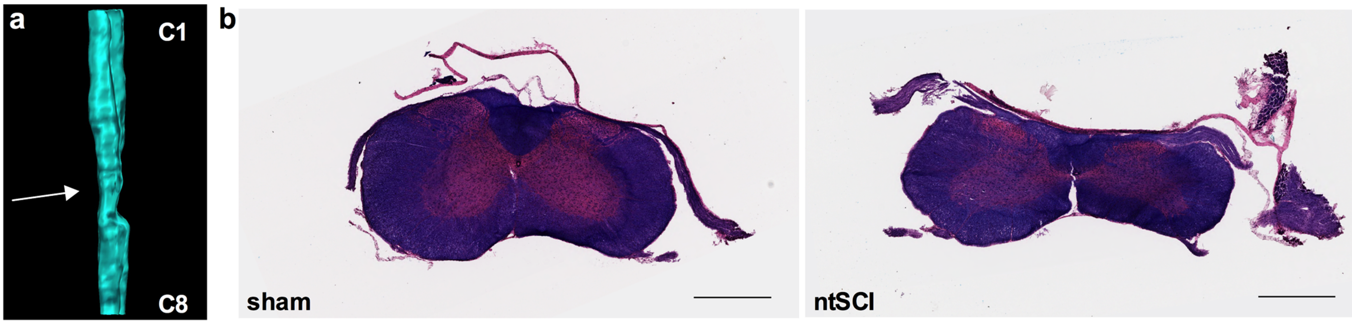
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

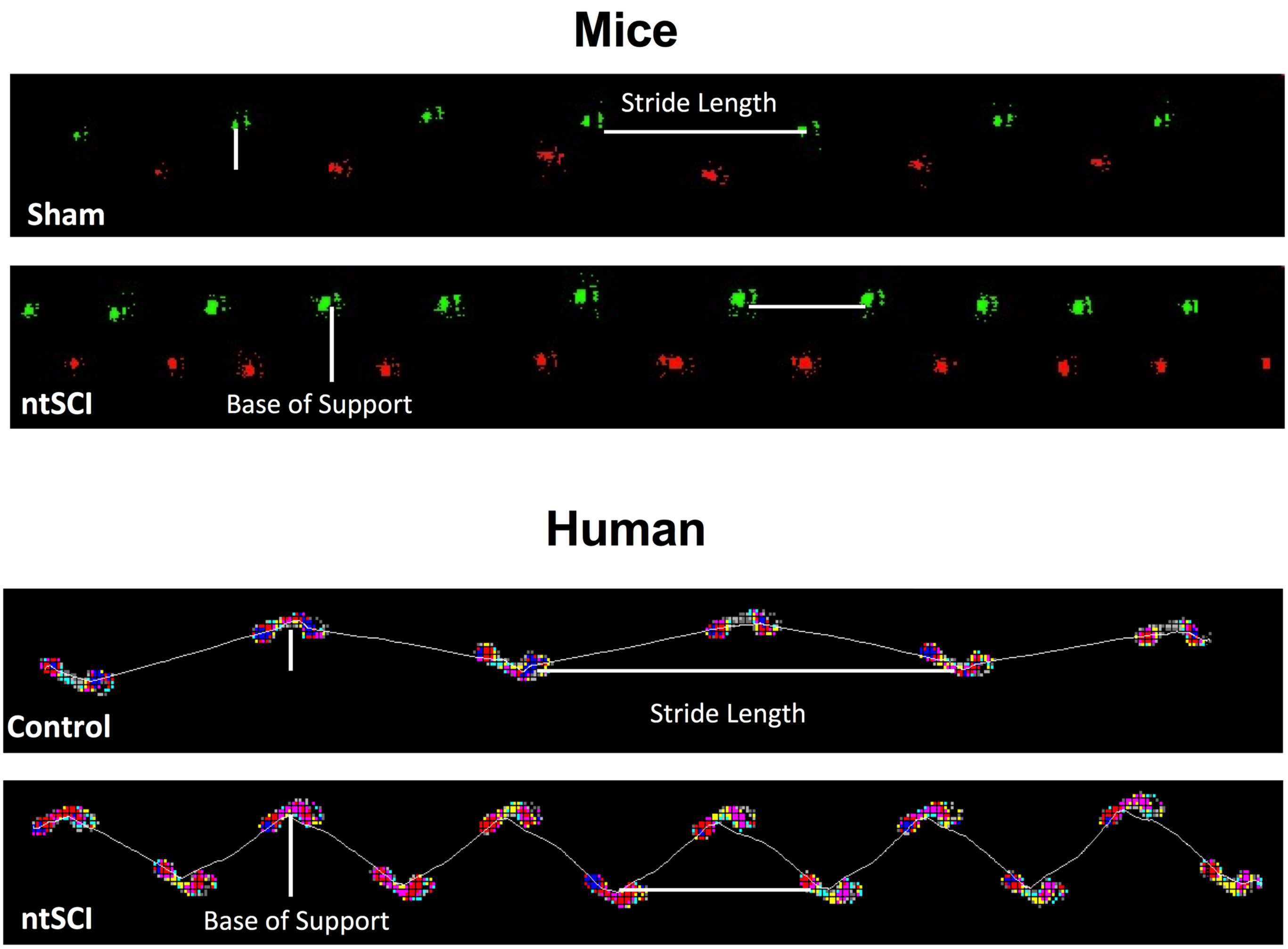
All quantification data from figures are provided in the Supplementary Information and all raw data are available upon reasonable request.

29. Satkunendrarajah, K. et al. Riluzole promotes motor and respiratory recovery associated with enhanced neuronal survival and function following high cervical spinal hemisection. *Exp. Neurol*. **276**, 59–71 (2016).
30. Klironomos, G. et al. New experimental rabbit animal model for cervical spondylotic myelopathy. *Spinal Cord* **49**, 1097–1102 (2011).
31. Mantilla, C. B., Zhan, W. Z. & Sieck, G. C. Retrograde labeling of phrenic motoneurons by intrapleural injection. *J. Neurosci. Methods* **182**, 244–249 (2009).
32. Bilney, B., Morris, M. & Webster, K. Concurrent related validity of the GAITRite walkway system for quantification of the spatial and temporal parameters of gait. *Gait Posture* **17**, 68–74 (2003).
33. Farr, T. D., Liu, L., Colwell, K. L., Whishaw, I. Q. & Metz, G. A. Bilateral alteration in stepping pattern after unilateral motor cortex injury: a new test strategy for analysis of skilled limb movements in neurological mouse models. *J. Neurosci. Methods* **153**, 104–113 (2006).
34. Whishaw, I. Q. & Pellis, S. M. The structure of skilled forelimb reaching in the rat: a proximally driven movement with a single distal rotatory component. *Behav. Brain Res*. **41**, 49–59 (1990).
35. Minic, Z. et al. Nanoconjugate-bound adenosine A_1 receptor antagonist enhances recovery of breathing following acute cervical spinal cord injury. *Exp. Neurol*. **292**, 56–62 (2017).
36. Montandon, G., Horner, R. L., Kinkead, R. & Bairam, A. Caffeine in the neonatal period induces long-lasting changes in sleep and breathing in adult rats. *J. Physiol. (Lond.)* **587**, 5493–5507 (2009).
37. McKay, L. C. & Feldman, J. L. Unilateral ablation of pre-Botzinger complex disrupts breathing during sleep but not wakefulness. *Am. J. Respir. Crit. Care Med*. **178**, 89–95 (2008).

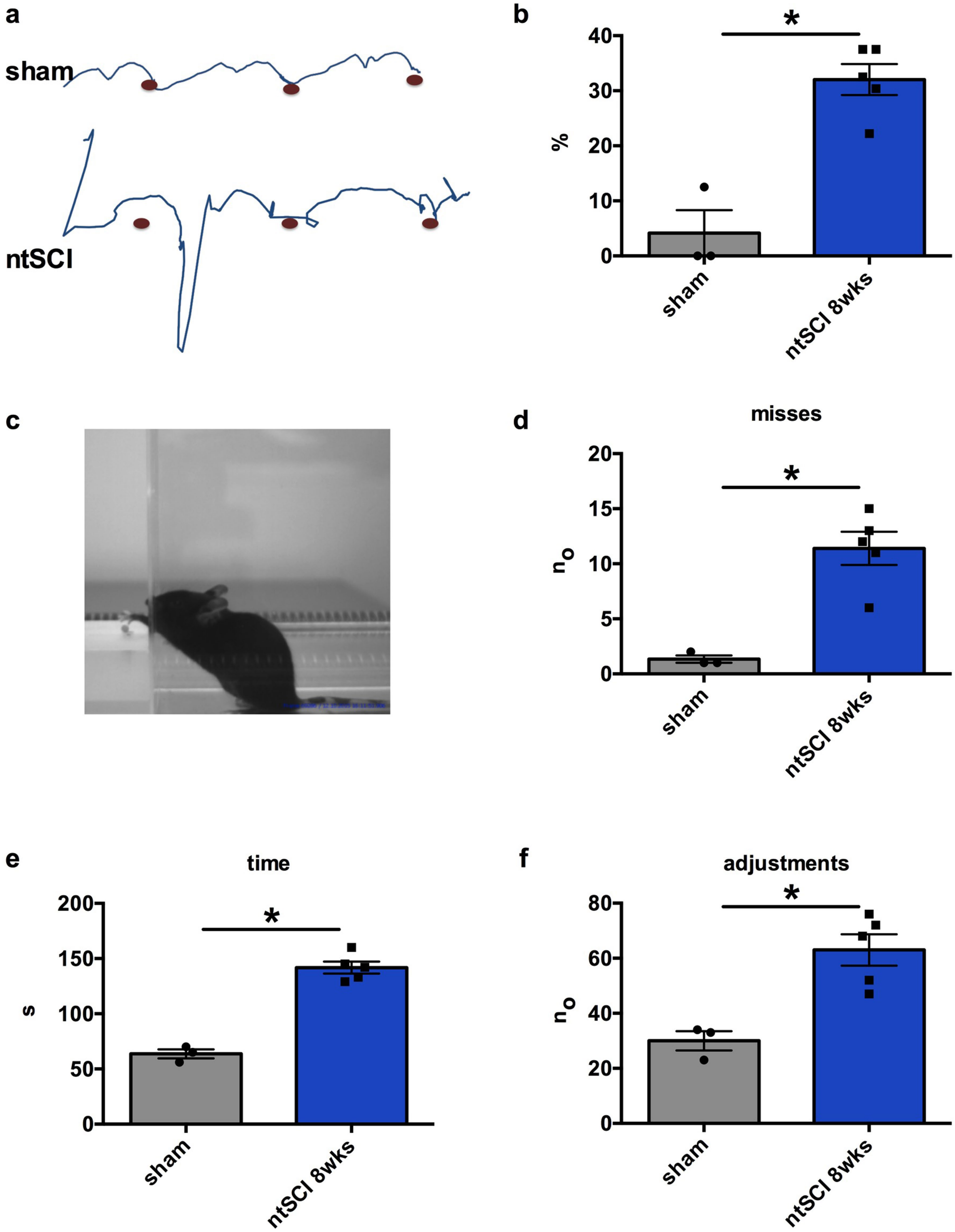
Extended Data Fig. 1 | MRI of the cervical spine of a patient with ntSCI due to degenerative cervical myelopathy. Arrows point to an area of severe compression of the cervical spinal cord at the C3–C5 levels, the area that houses the PMNs. Of note, this patient did not have clinical respiratory deficits.



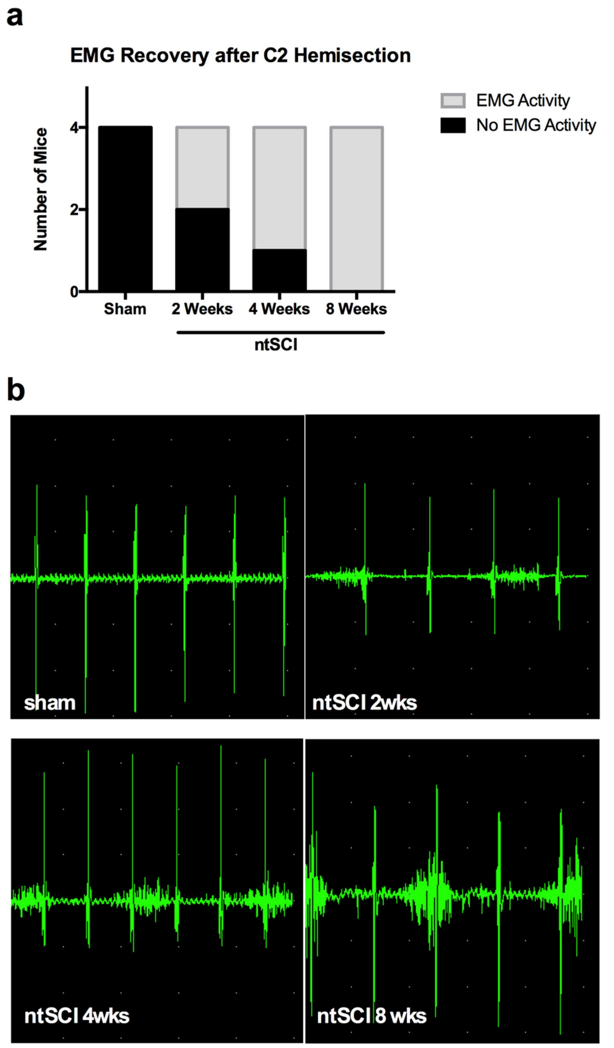
Extended Data Fig. 2 | ntSCI mice demonstrate notable compression of the cervical spinal cord. a, Sagittal 3D view of a reconstructed cervical spinal cord of a ntSCI mouse extending from C1 (rostral, r) to C8 (caudal, c) demonstrates the compressive injury at C4–C6 spinal level (arrow). **b**, Cervical spinal cord sections of ntSCI and sham mice stained with haematoxylin and eosin and luxol fast blue (representative of 4 biological replicates). Scale bar, 600 μm.



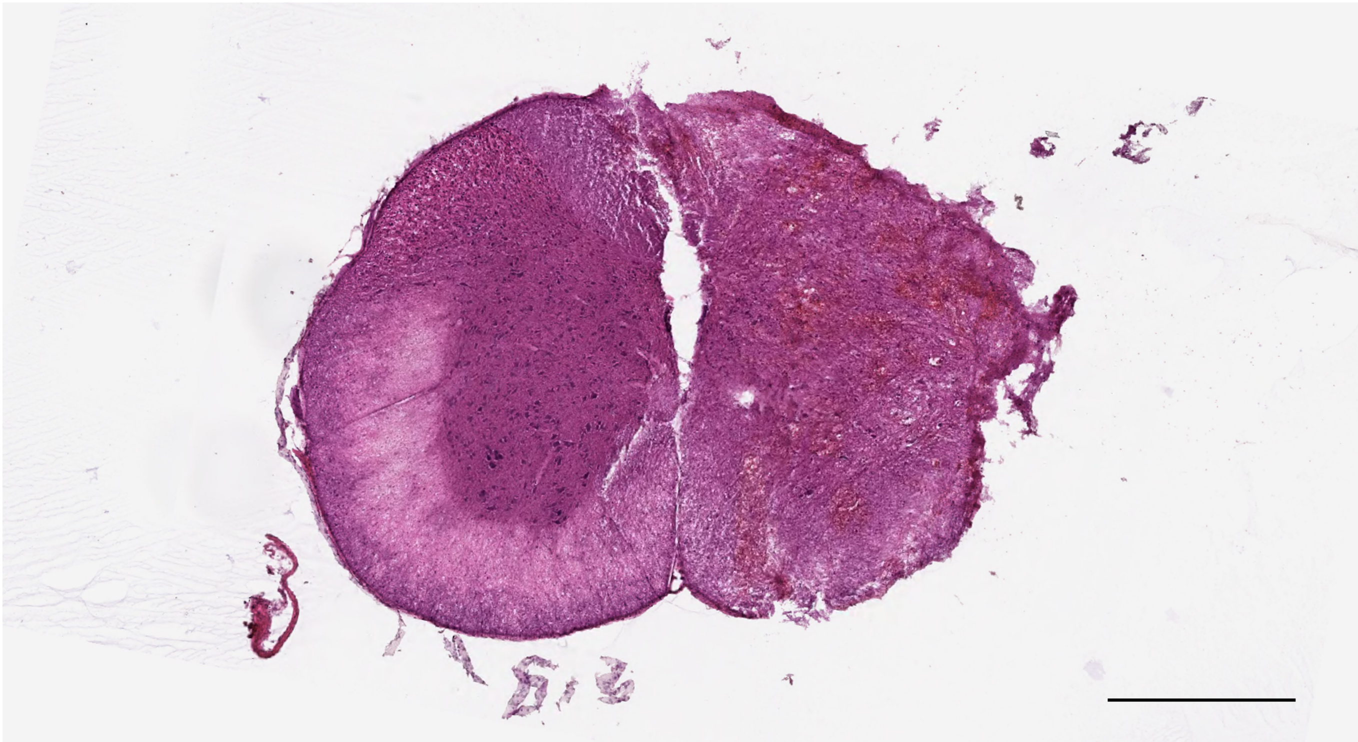
Extended Data Fig. 3 | Mice and humans with ntSCI display similar gait dysfunction. Representative footprint images from both a mouse (hindpaws only are displayed) and a human ntSCI subject demonstrating decreased stride length and increased base of support during walking relative to control subjects.



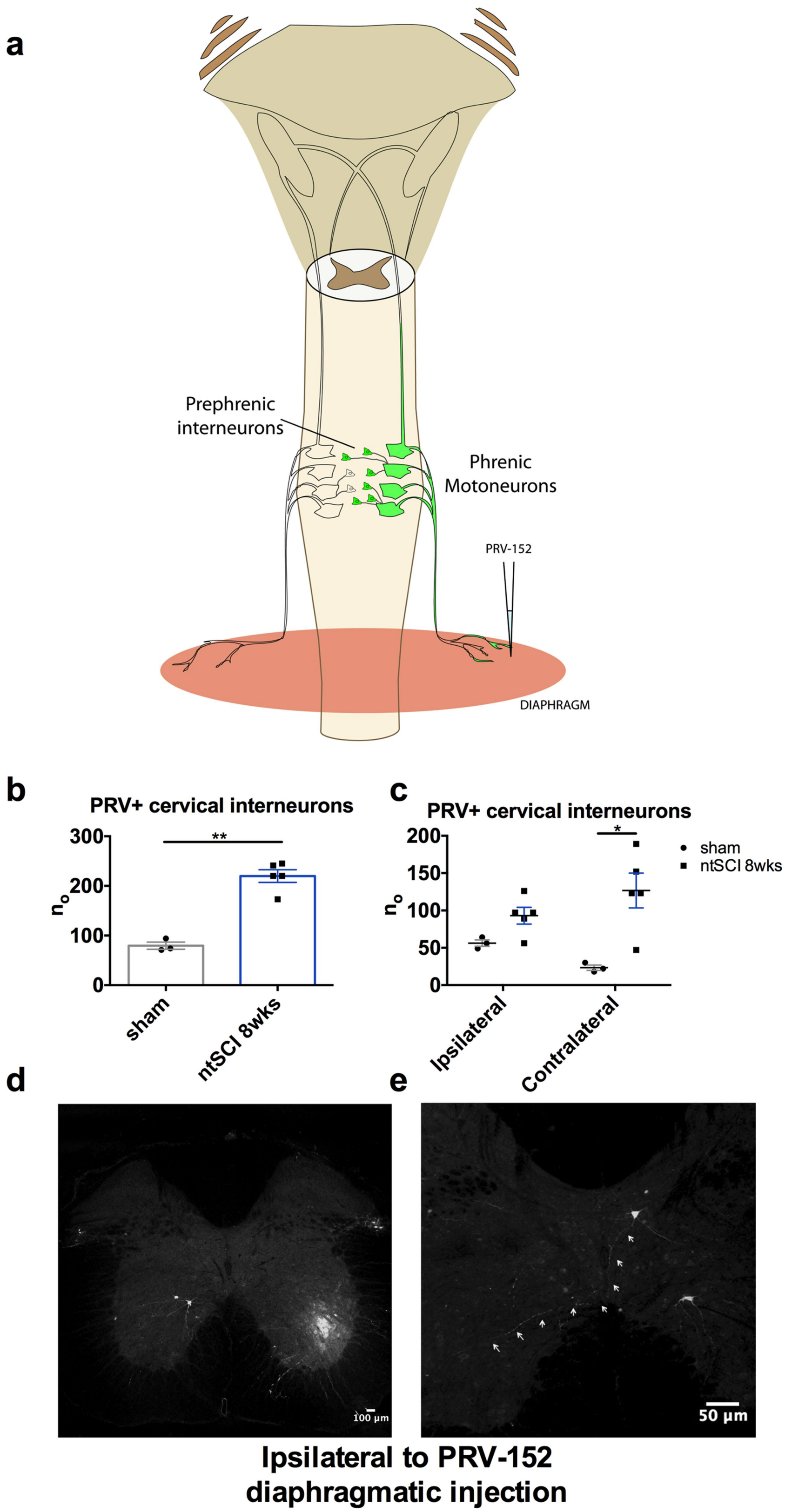
Extended Data Fig. 4 | ntSCI mice display a significant loss of manual dexterity. a, Representative position–time graph of forepaw movements during skilled walking on a horizontal ladder. **b**, Bar graph demonstrating an increased number of incorrect forepaw placements as a percentage of total placements in ntSCI mice during ladder walk ($P = 0.0012$, 95% CI = 16.02 to 39.69). **c**, Representative image showing reaching and grasping by a ntSCI mouse. **d**, The number of unsuccessful reaches was greater in ntSCI versus sham mice ($P = 0.0025$, 95% CI = 5.126 to 15.01). **e**, The time taken (seconds) to consume a whole length of pasta during the Capellini handling test was increased in ntSCI mice ($P = 5.71 \times 10^{-5}$, 95% CI = 59.06 to 97.21). **f**, The number of grip adjustments made by the mice during the consumption of the pasta ($P = 0.0063$, 95% CI = 13.34 to 52.66). Statistical significance was determined using an independent samples two-tailed *t*-test in **b**, **d**–**f**; $n = 3$ mice for sham and $n = 5$ mice for ntSCI. Data are mean ± s.e.m. $*P < 0.05$.



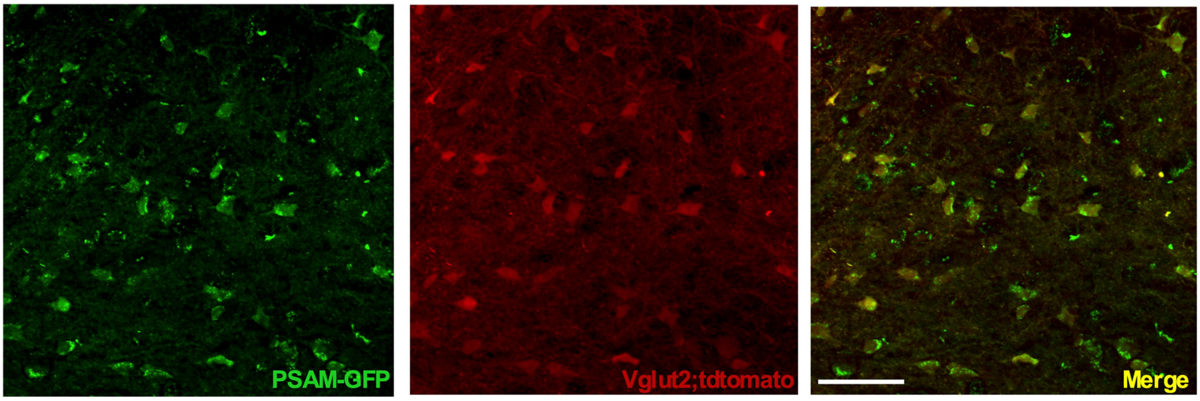
Extended Data Fig. 5 | Functional plasticity of the cervical respiratory network under ntSCI. a, Stacked column graph demonstrating the number of animals with recovered ipsilateral (left) hemidiaphragmatic electromyography (EMG) activity immediately after left C2 hemisection ($n = 4$ mice per group). Length of compression in the ntSCI animals was related to the proportion of animals with ipsilateral inspiratory activity after C2 hemisection ($P = 0.0034$, chi-square test for trend). **b**, Representative EMG recordings from the left hemidiaphragm immediately after left C2 hemisection in sham as well as in 2-, 4- and 8-week ntSCI mice.



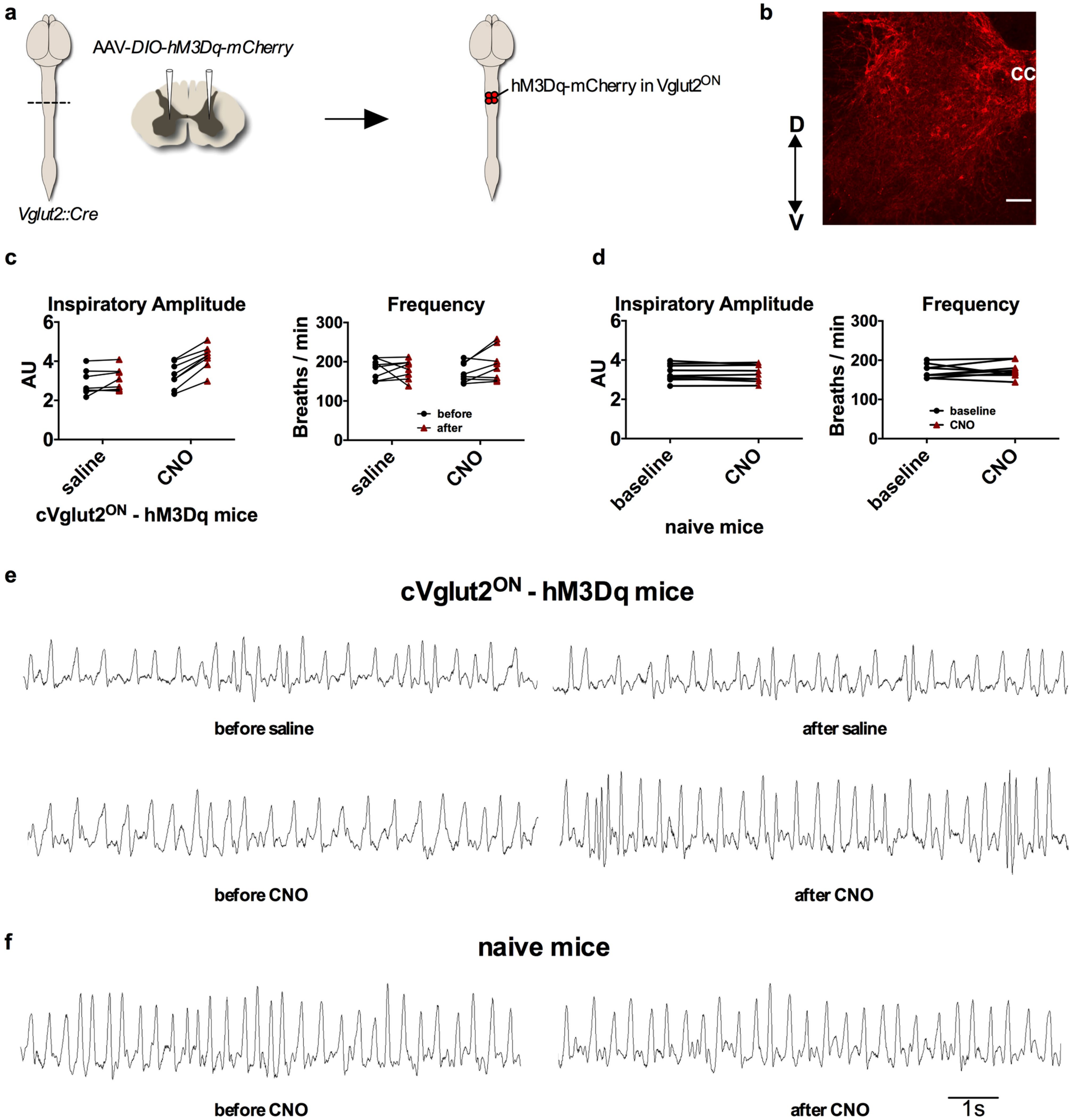
Extended Data Fig. 6 | C2 hemisection. Haematoxylin and eosin-stained cervical spinal cord section from a mouse subjected to a left C2 hemisection injury. Scale bar, 600 μm. Staining was performed on 22 separate C2 hemisection samples with similar results.



Extended Data Fig. 7 | Anatomical plasticity of the cervical respiratory network in ntSCI. a, Schematic demonstrating the injection paradigm to assess the anatomical plasticity of the cervical respiratory network after ntSCI. **b**, **c**, Bar graphs demonstrating a greater number of cervical interneurons connected to the preserved PMNs ($P = 2.0 \times 10^{-4}$, 95% CI = 96.44 to 183.8, unpaired *t*-test, two-tailed), and more specifically, a greater number of contralateral cervical interneurons in 8-week ntSCI mice ($P = 0.0024$, 95% CI = −166.2230 to −40.71032, repeated-measures ANOVA with Sidak's post hoc test for ipsilateral and contralateral comparisons; $n = 3$ mice for sham group and $n = 5$ mice for ntSCI). **d**, Confocal image of 30-µm-thick transverse cervical spinal cord section demonstrating PRV-traced PMNs on the left (ipsilateral) and prephrenic interneurons in the contralateral ventromedial region; the area where most prephrenic interneurons were located. **e**, Projections of the PRV152-traced pre-phrenic interneurons to the contralateral area housing the PMNs. Visualization of PRV152 was performed for all samples ($n = 3$ mice for sham and $n = 5$ mice for ntSCI) with similar results. Data are mean ± s.e.m. $^*P < 0.05$, $^{**}P < 0.001$.



Extended Data Fig. 8 | Expression of PSAM in midcervical excitatory neurons. High magnification confocal images demonstrate the expression of PSAM (GFP^+) in only the excitatory cells ($tdTomato^+$) of the intermediate grey matter. The images were taken from the same sections depicted in Fig. 2c, and are representative of results from 6 biological replicates. Scale bar, 50 μm.



Extended Data Fig. 9 | Stimulation of cervical eINs enhances respiratory output. **a**, Scheme demonstrating the approach used to express hM3Dq–mCherry in the eINs of the intermediate grey matter of C4–C5 spinal level. The Cre-dependent AAV-DIO-hM3Dq-mCherry viral construct was injected in the intermediate grey matter area of the cervical spinal cord of uninjured *Vglut2::cre* mice (VGLUT2-hM3Dq). **b**, Representative confocal image demonstrating the expression of hM3Dq–mCherry in the glutamatergic cells of the area of interest (based on 8 biological replicates). Note the lack of expression in the area of motor neurons. CC denotes the central canal. Scale bar, 100 μm. **c**, Respiratory measurements showed increased inspiratory amplitude but no change in respiratory frequency following stimulation of the cervical eINs ($P = 0.9042$, repeated-measures ANOVA; $n = 8$ mice for saline control and CNO groups). The interaction between the change in respiratory amplitude after injection, and treatment (saline versus CNO) was highly significant ($P = 0.0002$, repeated-measures ANOVA), indicating that the relative change after saline injection is not equal to the relative change after CNO ($n = 8$ mice for saline control and CNO groups). **d**, CNO did not change either the inspiratory amplitude ($P = 0.3513$, paired two-tailed t-test, 95% CI = −0.1077 to 0.04301, $n = 9$ mice) or the respiratory frequency ($P = 0.6906$, paired two-tailed t-test, 95% CI = −10.70 to 15.37, $n = 9$ mice) in *Vglut2::cre* mice lacking the AAV-DIO-hM3Dq-mCherry injections. **e**, Representative respiratory activity from VGLUT2-hM3Dq mice before and after saline or CNO administration. **f**, Representative respiratory activity from *Vglut2::cre* mice lacking AAV-DIO-hM3Dq-mCherry injection, before and after CNO administration.

Extended Data Table 1 | Statistics

Figure	Test	Group-wise Comparison	Statistic	Degrees of Freedom	Sample Size	Observed Effect Size*	p-Value
1a	t-test (two-tailed)		t=0.2740 t=0.7845 t=2.133	df=8 df=8 df=8	n=5/group n=5/group n=5/group	d_{CO2}=0.173 d_{pH}=1.36 d_{SaO2}=0.496	p=0.7910 p=0.4554 p=0.0655
1d	t-test (two-tailed)		t=4.165 t=0.550 t=0.5941	df=12 df=12 df=12	n=7/group n=7/group n=7/group	d_{area}=0.290 d_{amp}=2.226 d_{dur}=0.317	p=0.0013 p=0.5919 p=0.5635
1e	ANOVA/Tukey's Post Hoc	Sham v 4 wks Sham v 8 wks	q=9.977 q=11.40	df=12 df=12	n=4/group n=4/group	f=2.83	$p=2.54\times10^{-4}$ $p=6.86\times10^{-5}$
1f	ANOVA/Tukey's Post Hoc	Sham v 2 wks Sham v 4 wks Sham v 8 wks	q=4.534 q=10.01 q=21.89	df=12 df=12 df=12	n=4/group n=4/group n=4/group	f=4.30	p=0.0331 $p=2.50\times10^{-4}$ $p=5.56\times10^{-8}$
2e	RM-ANOVA	Interaction	F_{Hyp} =46.87 F_{Amp}=57.15	df= 1, 11	Sham= 7 ntSCI= 6	f=2.27	$p=2.78\times10^{-5}$ $p=1.11\times10^{-5}$
3d	Paired t-test (two-tailed)		t=19.66	df=9	n=10	dz=6.22	$p=1.06\times10^{-8}$
4e	Mann-Whitney-U test (two-tailed)		U=0		SCI+Saline= 8 SCI+CNO= 9	d=5.37 d=4.62	$p=8.23\times10^{-5}$
Ext. 4b	t-test (two-tailed)		t=5.761	df=6	Sham= 3 ntSCI= 5	d=4.11	p=0.0012
Ext. 4d	t-test (two-tailed)		t=4.986	df=6	Sham= 3 ntSCI= 5	d=4.17	p=0.0025
Ext. 4e	t-test (two-tailed)		t=10.02	df=6	Sham= 3 ntSCI= 5	d=7.89	$p=5.71\times10^{-5}$
Ext. 4f	t-test (two-tailed)		t=4.108	df=6	Sham= 3 ntSCI= 5	d=3.30	p=0.0063
Ext. 5a	Chi-Square test for trend		χ^2=8.584	df=1	n=4/group	ϕ_c=0.73	p=0.0034
Ext. 7b	t-test (two-tailed)		t=7.847	df=6	Sham= 3 ntSCI= 5	d=6.34	p=0.0002
Ext. 7c	RM-ANOVA/ Sidak's Post Hoc	Ipsilateral Contralateral	t=1.492 t=4.209	df=12	Sham= 3 ntSCI= 5	f=3.20	p=0.2971 p=0.0024
Ext. 9c	RM-ANOVA	Interaction Saline v CNO	F_{Amp}=26.27 F_{Freq}=0.015	df= 1, 14	Sham= 8 ntSCI= 8	f=1.37 f=0.03	p=0.0002 p=0.9042
Ext. 9d	Paired t-test (two-tailed)		t_{Amp}=0.9897 t_{Freq}=0.4128	df=8	n=9	dz=0.33 dz=0.14	p=0.3513 p=0.6906

*Effect size estimates: d = Cohen's d; dz = Cohen's dz (standardized difference score); f = Cohen's f, ϕ_c = Cramér's V.

https://doi.org/10.1038/s41586-018-0597-x

IRE1α–XBP1 controls T cell function in ovarian cancer by regulating mitochondrial activity

Minkyung Song[1,2,3], Tito A. Sandoval[2,3], Chang-Suk Chae[2,3], Sahil Chopra[1,2,3], Chen Tan[2,3], Melanie R. Rutkowski[4], Mahesh Raundhal[5,6], Ricardo A. Chaurio[7], Kyle K. Payne[7], Csaba Konrad[8], Sarah E. Bettigole[9], Hee Rae Shin[9], Michael J. P. Crowley[1], Juan P. Cerliani[10], Andrew V. Kossenkov[11], Ievgen Motorykin[12], Sheng Zhang[12], Giovanni Manfredi[8], Dmitriy Zamarin[13], Kevin Holcomb[2,3], Paulo C. Rodriguez[7], Gabriel A. Rabinovich[10,14], Jose R. Conejo-Garcia[7], Laurie H. Glimcher[5,6,15]* & Juan R. Cubillos-Ruiz[1,2,3,15]*

Tumours evade immune control by creating hostile microenvironments that perturb T cell metabolism and effector function[1–4]. However, it remains unclear how intra-tumoral T cells integrate and interpret metabolic stress signals. Here we report that ovarian cancer—an aggressive malignancy that is refractory to standard treatments and current immunotherapies[5–8]—induces endoplasmic reticulum stress and activates the IRE1α–XBP1 arm of the unfolded protein response[9,10] in T cells to control their mitochondrial respiration and anti-tumour function. In T cells isolated from specimens collected from patients with ovarian cancer, upregulation of *XBP1* was associated with decreased infiltration of T cells into tumours and with reduced *IFNG* mRNA expression. Malignant ascites fluid obtained from patients with ovarian cancer inhibited glucose uptake and caused *N*-linked protein glycosylation defects in T cells, which triggered IRE1α–XBP1 activation that suppressed mitochondrial activity and IFNγ production. Mechanistically, induction of XBP1 regulated the abundance of glutamine carriers and thus limited the influx of glutamine that is necessary to sustain mitochondrial respiration in T cells under glucose-deprived conditions. Restoring *N*-linked protein glycosylation, abrogating IRE1α–XBP1 activation or enforcing expression of glutamine transporters enhanced mitochondrial respiration in human T cells exposed to ovarian cancer ascites. XBP1-deficient T cells in the metastatic ovarian cancer milieu exhibited global transcriptional reprogramming and improved effector capacity. Accordingly, mice that bear ovarian cancer and lack XBP1 selectively in T cells demonstrate superior anti-tumour immunity, delayed malignant progression and increased overall survival. Controlling endoplasmic reticulum stress or targeting IRE1α–XBP1 signalling may help to restore the metabolic fitness and anti-tumour capacity of T cells in cancer hosts.

IRE1α excises a 26-nucleotide fragment from the *XBP1* mRNA, under endoplasmic reticulum (ER) stress, to generate a spliced version that encodes the functionally active XBP1s protein[9]. This transcription factor mediates adaptation to ER stress by inducing genes that are involved in protein folding and quality control[10]. IRE1α–XBP1 endows malignant cells with tumorigenic capacity[11], while subverting the function of cancer-associated myeloid cells[12–14]. However, it remains unknown whether this pathway operates intrinsically in T cells to influence malignant progression.

Intra-tumoral and ascites-resident $CD4^+$ and $CD8^+$ T cells isolated from specimens derived from patients with ovarian cancer demonstrated increased *XBP1* mRNA splicing, compared with peripheral T cells from women who are free of cancer (Fig. 1a, b). *XBP1* levels in intra-tumoral T cells were correlated with expression of *HSPA5* and *DDIT3*, which are genes that indicate activation of the unfolded protein response (UPR); this suggests that these cells experience ER stress in situ (Fig. 1c). Increased expression of *XBP1*, *HSPA5* and *DDIT3* was associated with reduced T cell infiltration in the specimens that we analysed (Fig. 1d). However, only *XBP1* expression correlated with decreased *IFNG* levels in intra-tumoral T cells (Fig. 1e), which suggests that IRE1α–XBP1 activation driven by ER stress might influence T cell functions in ovarian cancer.

XBP1 splicing was observed mainly in T cells present in ovarian cancer ascites (Fig. 1b), which is an immunomodulatory and tumorigenic fluid that often accumulates in patients with metastatic or recurrent disease[6,15]. We exploited this milieu to examine whether ovarian cancer induces IRE1α–XBP1 in T cells to control their activity. We focused on $CD4^+$ T cells because they are the predominant leukocyte population in ovarian cancer ascites[16–19], and because the mechanisms regulating their protective capacity in this setting remain unclear. Pre-activated $CD4^+$ T cells from women who are free of cancer exhibited a dose-dependent increase in *XBP1* upon treatment with cell-free supernatants of ascites from patients with ovarian cancer (Extended Data Fig. 1a). Analyses based on fluorescence-activated cell sorting (FACS) confirmed the induction of XBP1s in response to exposure to ascites (Fig. 2a, b). T cells treated with the ER stressor tunicamycin demonstrated strong XBP1s staining that was abrogated by the IRE1α inhibitor 4μ8C (Extended Data Fig. 1b), validating the specificity of XBP1s detection by FACS. Hypoxia, acidic pH and nutrient deprivation disrupt ER homeostasis and trigger the UPR[11]. Although ovarian cancer ascites is hypoxic in vivo[20], we observed XBP1s induction in T cells exposed to this fluid even under normoxia (Fig. 2a, b). The pH of independent ascites samples was within neutral range. Ascites treatment modestly augmented the levels of reactive oxygen species, but this effect was not responsible for *XBP1* induction in T cells (Extended Data Fig. 1c, d). Glucose is essential for *N*-linked protein glycosylation and, therefore, alterations in its availability can provoke ER stress[21]. No correlation was found between the ascites glucose concentration and *XBP1* induction in $CD4^+$ T cells (Extended Data Fig. 1e, f). However, ascites exposure suppressed expression of the major glucose transporter GLUT1 in $CD4^+$ T cells (Fig. 2c, d). This result was consistent with the observation that T cells present in the ascites of patients with ovarian cancer demonstrate negligible GLUT1 surface expression

[1]Weill Cornell Graduate School of Medical Sciences, New York, NY, USA. [2]Department of Obstetrics and Gynecology, Weill Cornell Medicine, New York, NY, USA. [3]Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [4]Department of Microbiology, Immunology and Cancer Biology, University of Virginia, Charlottesville, VA, USA. [5]Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA, USA. [6]Department of Medicine, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, USA. [7]Department of Immunology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA. [8]Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA. [9]Quentis Therapeutics, Inc, New York, NY, USA. [10]Laboratorio de Inmunopatología, Instituto de Biología y Medicina Experimental (IBYME), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina. [11]Center for Systems and Computational Biology, The Wistar Institute, Philadelphia, PA, USA. [12]Proteomics & Mass Spectrometry Facility, Institute of Biotechnology, Cornell University, Ithaca, NY, USA. [13]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [14]Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina. [15]These authors jointly supervised this work: Laurie H. Glimcher, Juan R. Cubillos-Ruiz. *e-mail: laurie_glimcher@dfci.harvard.edu; jur2016@med.cornell.edu

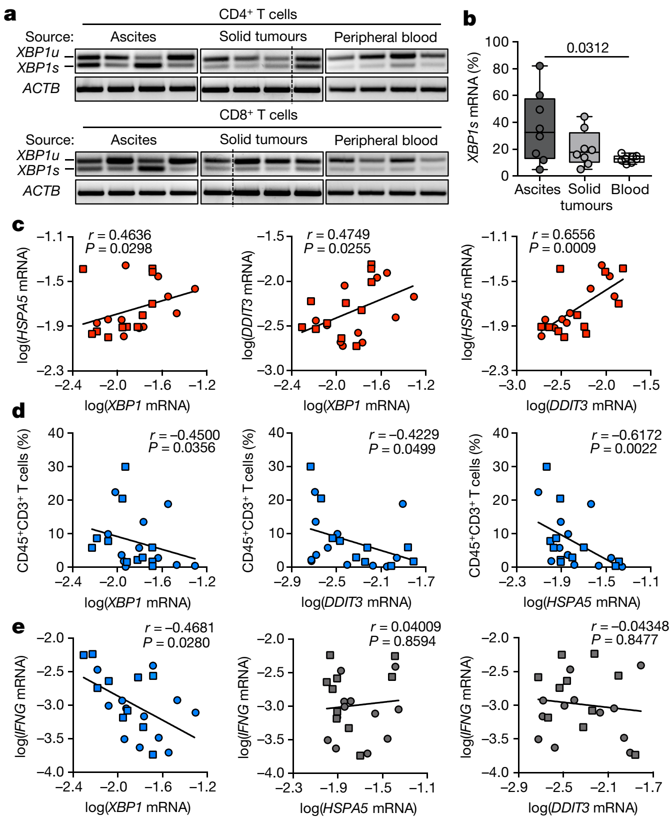


Fig. 1 | IRE1α–XBP1 activation in human T cells that infiltrate ovarian cancer. a, *XBP1* splicing assays for CD4$^+$ or CD8$^+$ T cells isolated from ascites or solid tumours of patients with ovarian cancer, or from blood of cancer-free female donors. *XBP1s*, spliced form; *XBP1u*, unspliced form. Data were generated from three independent experiments. **b**, Frequency of spliced *XBP1* divided by total *XBP1* in T cells sorted from the indicated sources ($n = 8$ per group). **c–e**, Pairwise analyses for sorted ovarian cancer-associated CD4$^+$ (circles) and CD8$^+$ (squares) T cells ($n = 22$ total). **c**, Expression of genes linked with ER stress responses. **d**, Proportion of CD45$^+$CD3$^+$ T cells present in the ovarian cancer specimen versus expression of the indicated genes in T cells isolated from the matched specimen analysed. **e**, *IFNG* versus genes associated with ER stress responses, in each specimen analysed. *n* values correspond to biologically independent samples (**b**–**e**). One-way ANOVA with Tukey's post-test, boxes represent median ± interquartile range and whiskers indicate minimum and maximum (**b**); Spearman's rank correlation test, Spearman coefficient (r) with P value (two-tailed), 95% confidence intervals for all correlation analyses (**c**–**e**) are described in Supplementary Table 2.

(Extended Data Fig. 1g). Glucose uptake was therefore compromised in ascites-exposed CD4$^+$ T cells, and this defect was associated with enhanced expression of *XBP1* mRNA and of XBP1s (Fig. 2e, Extended Data Fig. 1h).

Protein glycosylation requires uridine diphosphate *N*-acetylglucosamine (UDP-GlcNAc), which is generated through the hexosamine biosynthetic pathway using glucose as a substrate[21]. Ascites-exposed CD4$^+$ T cells displayed *N*-linked protein glycosylation defects (Extended Data Fig. 1i, Extended Data Table 1), but supplementation with *N*-acetylglucosamine (GlcNAc)—which serves as direct substrate for *N*-linked glycosylation—attenuated XBP1 activation in this setting (Extended Data Fig. 1j). Limited glucose uptake in this setting also dampened the glycolytic capacity of CD4$^+$ T cells (Fig. 2f). The oxygen consumption rates of ascites-exposed CD4$^+$ T cells decreased in a dose-dependent manner (Fig. 2g, Extended Data Fig. 1k), which suggests that this fluid disturbs mitochondrial activity. T cell mitochondrial respiration also diminished when protein *N*-linked glycosylation was directly inhibited using tunicamycin (Extended Data Fig. 1l, m). Because pyruvate was always present in the culture medium, the mitochondrial perturbations observed were unlikely to be caused by reduced T cell glycolysis in the presence of ascites. Therefore, IRE1α–XBP1 activation might directly affect T cell mitochondrial function. Blocking IRE1α with 4μ8C inhibited induction of XBP1s (Fig. 2h) while enhancing mitochondrial respiration in human CD4$^+$ T cells exposed to ovarian cancer ascites (Fig. 2i, Extended Data Fig. 1n). IRE1α inhibition did not modulate glucose uptake or glycolysis in ascites-treated T cells (Extended Data Fig. 1o), which confirms that impaired glucose import drives IRE1α–XBP1 activation. Attenuating XBP1s induction using GlcNAc (Extended Data Fig. 1j) also augmented mitochondrial respiration in T cells treated with supernatants of ovarian cancer ascites (Fig. 2j, Extended Data Fig. 1p, q). Thus, reduced glucose import by ascites-exposed T cells not only dampens glycolysis but also impairs *N*-linked protein glycosylation, leading to ER stress and mitochondrial dysfunction driven by IRE1α–XBP1.

We generated conditional knockout mice (*Xbp1*$^{f/f}$*Cd4*^{cre}) to understand how XBP1 regulates T cell function (Extended Data Fig. 2a). The absolute cell numbers in lymphoid tissues—as well as the frequency of thymocytes, peripheral CD4$^+$ and CD8$^+$ T cells, and their corresponding subsets—were unaltered upon ablation of XBP1 (Extended Data Fig. 2b–j). Loss of XBP1 in T cells did not affect the proportions of other immune-cell populations in lymphoid tissues (Extended Data Fig. 2k, l). Mixed bone marrow chimaeras, using wild-type mice and mice that lack XBP1 in the entire haematopoietic compartment, demonstrated normal T cell reconstitution (Extended Data Fig. 2m). Proliferation and cell-cycle progression was unaltered in CD4$^+$ T cells that lack XBP1 (Extended Data Fig. 2n, o). XBP1 therefore appears to be dispensable for T cell development and baseline functions in naive mice.

Glucose starvation for 6 h triggered *Xbp1* splicing and induction of gene markers for the UPR (Extended Data Fig. 3a) in CD4$^+$ T cells that had previously been activated through CD3 and CD28. Glutamine withdrawal alone did not induce IRE1α–XBP1 at the time point we analysed, which suggests that T cells primarily activate this pathway in response to glucose limitation. Mitochondrial respiration was comparable in wild-type and XBP1-deficient CD4$^+$ T cells when glucose was present (Fig. 3a). The maximal oxygen consumption rate in wild-type CD4$^+$ T cells decreased by about 50% upon glucose withdrawal but XBP1-deficient CD4$^+$ T cells demonstrated superior mitochondrial respiratory capacity in this condition (Fig. 3a), which supports our observations using IRE1α inhibition (Fig. 2i) or GlcNAc supplementation in ascites-exposed human CD4$^+$ T cells (Fig. 2j). T cell mitochondrial structure, morphology or mass were not affected by XBP1 deficiency, and expression of PGC1α also remained unaltered even upon glucose withdrawal or glycolysis inhibition (Extended Data Fig. 2p–r). Thus, XBP1 might restrain the use of alternative carbon sources that support mitochondrial respiration in the absence of glucose.

Glutamine metabolism maintains the tricarboxylic acid (TCA) cycle in response to glucose deprivation or reduced pyruvate supply[22]. Blocking mitochondrial glutamine use in glucose-deprived CD4$^+$ T cells further reduced their maximal oxygen consumption rate (Extended Data Fig. 3b). Because glutamine consumption in haematopoietic cells depends on glucose availability[23], we hypothesized that XBP1 activation under glucose restriction could affect this process. Glucose starvation impaired [U-^{13}C]glutamine uptake in wild-type CD4$^+$ T cells, but XBP1-deficient T cells in the same condition exhibited superior glutamine import and use, as evidenced by increased ^{13}C enrichment in glutamine, glutamate and various TCA-cycle metabolites (Extended Data Fig. 3c–i). Consequently, glucose-deprived T cells that lack XBP1 maintained higher levels of total mitochondrial metabolite pools derived from [U-^{13}C]glutamine compared with their wild-type counterparts (Fig. 3b). XBP1 therefore restricts glutamine influx in T cells under glucose starvation.

IRE1α–XBP1 activation can mediate the degradation of glutamine transporters under ER stress by activating the ER-associated degradation system[24–26], probably as a strategy to dampen metabolism under adverse conditions. Glucose withdrawal decreased the protein levels of the glutamine transporters ASCT2, SNAT1 and SNAT2 in wild-type

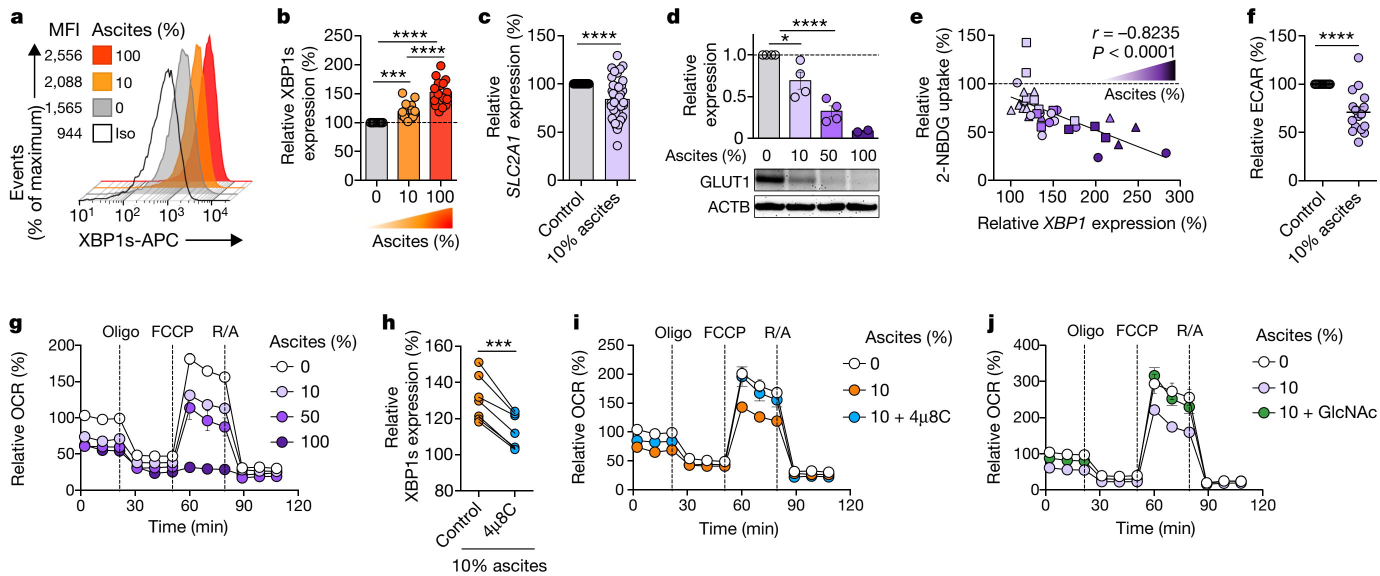


Fig. 2 | Ovarian cancer ascites limits glucose uptake and causes IRE1α–XBP1-mediated mitochondrial dysfunction in human CD4^{+} T cells. **a–g**, T cells were activated by CD3 and CD28 stimulation for 16 h in the absence or presence of supernatants from ovarian cancer ascites, at the indicated concentrations. **a**, **b**, Histograms (**a**) and quantification (**b**) of XBP1s staining ($n = 16$); iso, isotype control; MFI, mean fluorescence intensity. **c**, *SLC2A1* expression was determined by quantitative PCR (qPCR) ($n = 48$). **d**, Immunoblot and quantification of GLUT1 in ascites-exposed CD4^{+} T cells. Density of GLUT1 was normalized to ACTB, and data are shown as the relative expression compared with the untreated control ($n = 4$ for 10% and 50% ascites; $n = 2$ for 100% ascites, all from two independent experiments). **e**, Glucose uptake was assessed using 2-NBDG and *XBP1* expression was determined in the same sample. Symbols depict ascites from three independent patients tested at increasing concentrations on CD4^{+} T cells from multiple donors ($n = 37$). **f**, **g**, Baseline extracellular acidification rate (ECAR) (**f**) and oxygen consumption rate (OCR) profile (**g**) of CD4^{+} T cells exposed to ascites ($n = 16$). **h–j**, CD4^{+} T cells were treated with 4μ8C (**h**, **i**) or GlcNAc (**j**) for 1 h and then stimulated by CD3 and CD28 for 16 h in the presence of 10% ascites. **h**, XBP1s determined by FACS ($n = 7$). **i**, OCR profile in 4μ8C-treated T cells exposed to ascites ($n = 9$). **j**, OCR for GlcNAc-treated T cells exposed to ascites ($n = 5$). Data are shown as mean ± s.e.m. (**b–d**, **f**, **g**, **i**, **j**). *n* values represent biologically independent samples (**b–j**). One-way ANOVA with Tukey's post-test (**b**); two-tailed Student's *t*-test (**c**, **f**); one-way ANOVA with Bonferroni's post-test (**d**). Spearman's rank correlation test, 95% confidence intervals −0.9076 to −0.6760 (**e**); two-tailed paired Student's *t*-test (**h**). $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$. FCCP, carbonyl cyanide-*p*-(trifluoromethoxy) phenylhydrazone; oligo, oligomycin; R/A, rotenone and antimycin.

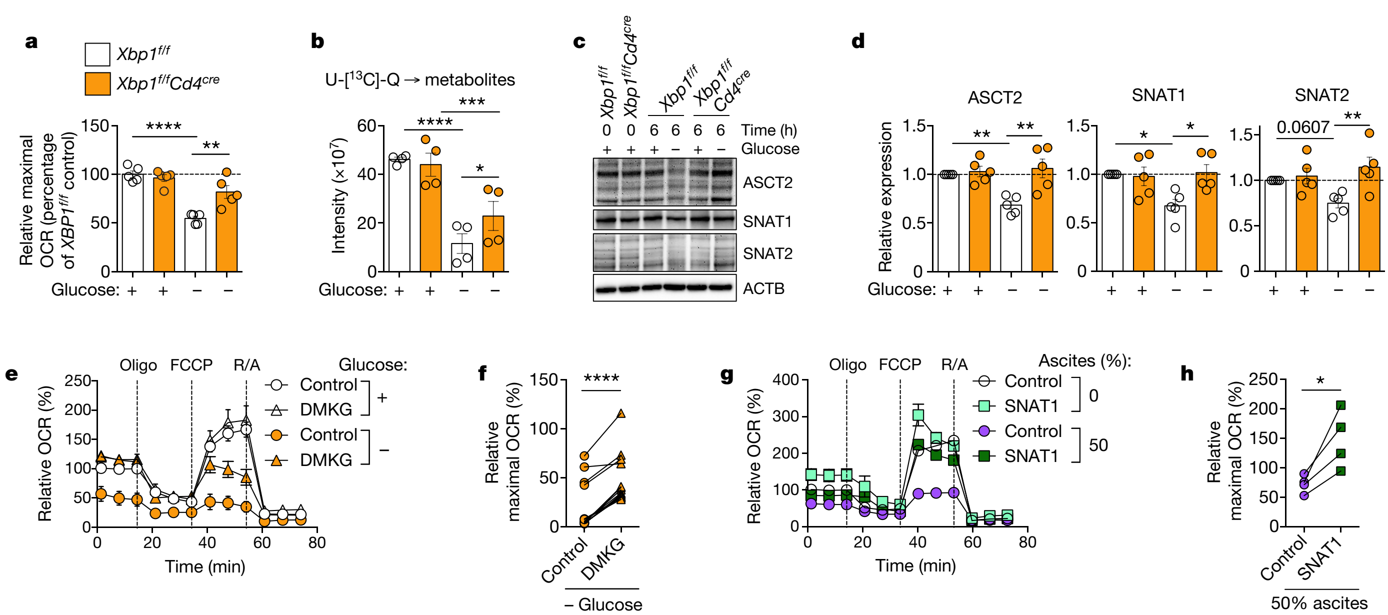


Fig. 3 | XBP1 limits glutamine influx in glucose-deprived CD4^{+} T cells. **a**, **b**, **d**, Naive splenic CD4^{+} T cells isolated from wild-type (white) or XBP1-deficient (orange) mice were activated by CD3 and CD28 stimulation for 48 h and then incubated for 6 h in the indicated medium. **a**, Maximal OCR of T cells in the presence or absence of glucose ($n = 5$). **b**, Glutamine tracing was performed as described in Methods, and relative abundance of total ^{13}C-labelled metabolites was determined ($n = 4$). U-[^{13}C]-Q, uniformly-labelled L-glutamine with ^{13}C. **c**, **d**, Immunoblot (**c**) and quantification (**d**) of glutamine transporters in the indicated T cells ($n = 5$ total from 5 independent experiments). **e**, **f**, OCR profile (**e**) and maximal OCR (**f**) in wild-type T cells treated with DMKG ($n = 14$). Data are presented as relative expression compared with wild-type T cells incubated in the presence of glucose (**a**, **d**, **f**). **g**, **h**, OCR profile (**g**) and maximal OCR (**h**) for SNAT1-overexpressing human CD4^{+} T cells exposed to ovarian cancer ascites ($n = 4$). Data are shown as relative expression compared with control-virus-transduced T cells that were not exposed to ascites. *n* values represent biologically independent samples (**a**, **b**, **d**, **f**, **h**). Data are shown as mean ± s.e.m. One-way ANOVA with Tukey's post-test (**a**, **b**, **d**); two-tailed paired Student's *t*-test (**f**, **h**). $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$.

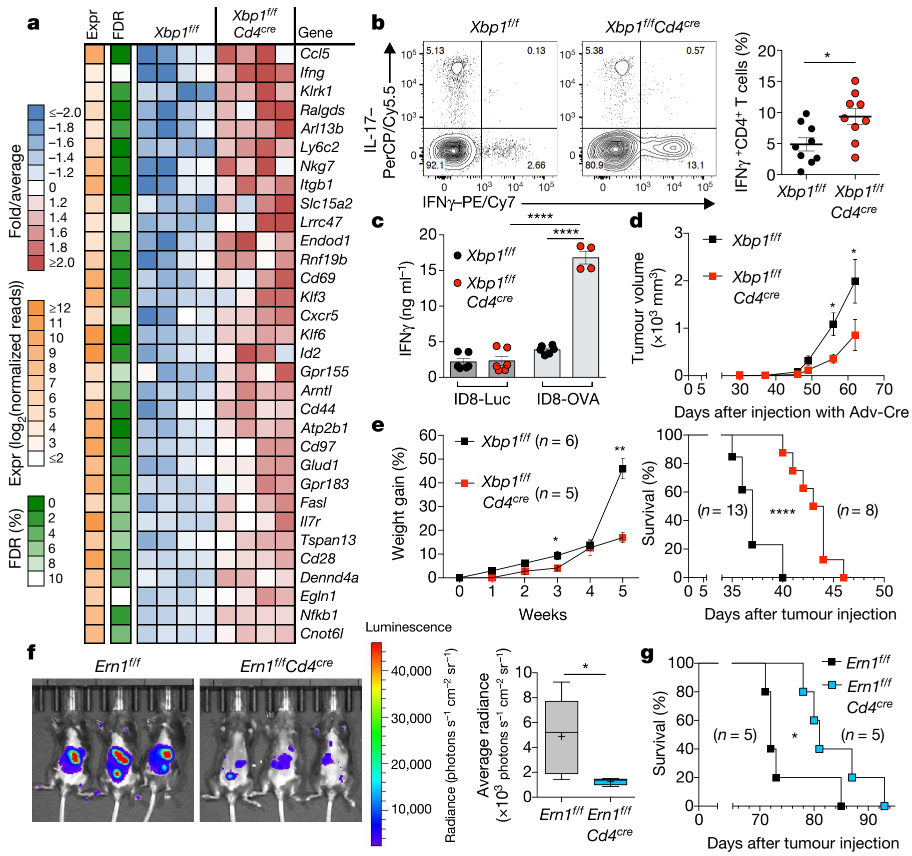


Fig. 4 | IRE1α–XBP1 signalling that is intrinsic to T cells promotes ovarian cancer progression. a, Transcriptional profiling of wild-type versus XBP1-deficient CD4$^+$ T cells sorted from the peritoneal cavity of mice that bore metastatic ID8-*Defb29*/*Vegfa* ovarian cancer for 20 days. Top-upregulated genes in XBP1-deficient CD4$^+$ T cells are shown ($n = 4$). FDR, false discovery rate; fold/average, fold change of expression relative to average normalized reads of all samples; Expr, $\log_2$ value of normalized reads. **b**, FACS analyses of intra-tumoral CD4$^+$ T cells from the indicated mice that bore metastatic ovarian cancer for 20–23 days. Representative intracellular staining for IFNγ and IL-17 (left) and global IFNγ analysis (right) in CD45$^+$CD3$^+$CD4$^+$ T cells ($n = 9$). **c**, IFNγ secretion by CD4$^+$ T cells isolated from the peritoneal cavity of the indicated ovarian-cancer-bearing mice upon ex vivo stimulation with OVA peptide ($n = 6$ for all groups, except for XBP1-deficient hosts challenged with ID8-OVA, $n = 4$). ID8-Luc, parental ID8 cells that express luciferase; ID8-OVA, parental ID8 cells that express ovalbumin protein. **d**, Growth of ovarian tumours driven by p53 loss and mutant KRAS in hosts reconstituted with bone marrow from the indicated genotypes ($n = 6$ for each genotype). ADV-Cre, adenovirus expressing functional Cre recombinase. **e**, Ascites accumulation (left) and overall survival (right) for the indicated mice bearing ID8-*Defb29*/*Vegfa* ovarian cancer. **f**, Imaging (left) and quantification (right) of peritoneal carcinomatosis in *Ern1*$^{f/f}$ or *Ern1*$^{f/f}$*Cd4*^{cre} mice that bore luciferase-expressing ID8 (ID8-Luc) ovarian cancer for 20 days ($n = 5$ for each genotype). **g**, Survival rates for mice depicted in **f**. n values represent biologically independent mice (**a**–**g**). Data are shown as mean ± s.e.m. (**b**–**e**). Boxes represent median ± interquartile range and whiskers indicate minimum and maximum (**f**). Two-tailed Student's t-test (**b**); one-way ANOVA with Tukey's post-test (**c**); two-tailed Mann–Whitney U-test (**d**–**f**); log-rank test (**e**, **g**). $*P < 0.05$, $**P < 0.01$, $****P < 0.0001$.

CD4$^+$ T cells (Fig. 3c, d). However, their XBP1-deficient counterparts demonstrated higher transporter expression under the same conditions (Fig. 3c, d), which was confirmed by immunofluorescence confocal microscopy (Extended Data Fig. 4a, b). Loss of XBP1 did not modulate expression of genes encoding glutamine carriers under glucose limitation (Extended Data Fig. 4c), but treatment with the proteasome inhibitor MG132 restored SNAT1 levels in wild-type T cells that were experiencing ER stress owing to glucose limitation (Extended Data Fig. 4d, e). These data suggest that XBP1 activation restricts glutamine influx in glucose-deprived CD4$^+$ T cells by controlling the abundance of glutamine carriers, probably through post-translational regulatory mechanisms.

Cells catabolize glutamine into α-ketoglutarate to replenish TCA-cycle intermediates and sustain mitochondrial respiration[22]. Treatment with dimethyl-α-ketoglutarate (DMKG), a cell-permeable analogue of α-ketoglutarate, improved maximal mitochondrial respiration in T cells under glucose restriction (Fig. 3e, f). Ascites-exposed human CD4$^+$ T cells—unable to import glucose and demonstrating mitochondrial dysfunction driven by IRE1α–XBP1 (Fig. 2)—also showed reduced levels of glutamine transporters (Extended Data Fig. 5a, b), but DMKG treatment was sufficient to increase their maximal respiratory capacity in this milieu (Extended Data Fig. 5c, d). Furthermore, overexpressing SNAT1 (Extended Data Fig. 5e, f) markedly enhanced mitochondrial function in human CD4$^+$ T cells exposed to ascites derived from patients with ovarian cancer (Fig. 3g, h). These results indicate that glucose restriction causes T cell mitochondrial dysfunction through XBP1-mediated inhibition of glutamine influx, and that overexpressing glutamine transporters or supplementing glutamine-derived TCA intermediates—such as α-ketoglutarate—can restore mitochondrial respiration in ascites-exposed T cells.

To evaluate the role of this pathway in vivo, we developed orthotopic ovarian cancer in female mice that lack XBP1 in T cells. CD4$^+$ and CD8$^+$ T cells isolated from ovarian tumours driven by p53 loss and mutant KRAS[27]—or from ascites and spleens of host mice bearing metastatic ovarian cancer that overexpresses *Defb29* and *Vegfa* in the ID8 cell line (hereafter, 'ID8-*Defb29*/*Vegfa*')[28]—demonstrated marked overexpression of total and spliced *Xbp1* transcripts, as well as robust induction of gene markers for the UPR (Extended Data Fig. 6a, b). However, T cells associated with ovarian cancer did not present signs of regulated IRE1α-dependent mRNA decay (Extended Data Fig. 6c, d).

Transcriptomic analyses, performed on CD4$^+$ T cells residing in the ascites of mice bearing metastatic ID8-*Defb29*/*Vegfa* ovarian cancer[28],

identified 151 differentially expressed genes in wild-type versus XBP1-deficient $CD4^+$ T cells. In the ascites, genes related to T cell activation—such as *Cd69*, *Cd44*, *Cd28* and *Nfkb1*—and mediators of anti-tumour immunity—such as *Ccl5*, *Ifng*, *Klrk1* and *Fasl*—were upregulated in XBP1-deficient $CD4^+$ T cells compared with their wild-type counterparts (Fig. 4a). These genes were not differentially expressed in wild-type versus XBP1-deficient splenic $CD44^{high}CD62L^{low}CD4^+$ T cells sorted from naive mice (Extended Data Fig. 7a), which demonstrates a function for XBP1 that is context-specific in T cells associated with ovarian cancer. Loss of XBP1 did not cause regulated IRE1α-dependent mRNA decay[29], nor did it alter the expression of transcription factors that govern helper T cell differentiation or $CD4^+$ regulatory T cell proportions at tumour sites (Extended Data Fig. 7b–d). Immunosuppressive gene signatures driven by TGFβ1, PGE_2 and IL-10 signalling were predicted to be repressed in XBP1-deficient $CD4^+$ T cells present in ovarian cancer ascites, whereas immuno-activating gene networks induced by ETS1, CD40 ligand engagement and protein kinases (such as RIPK2 and MAP3K14) were predicted to be activated in these cells (Extended Data Fig. 7e). Accordingly, downstream cellular functions such as activation, proliferation and migration of immune cells were predicted to be induced in ascites-infiltrating $CD4^+$ T cells that lack XBP1 (Extended Data Fig. 7f). Mice lacking XBP1 in T cells that bore ovarian cancer had increased proportions of $CD4^+$ T cells that secrete IFNγ at tumour sites, compared with their wild-type counterparts (Fig. 4b, Extended Data Fig. 7g). However, the frequency of infiltrating $CD4^+$ T cells that produce IL-17 remained unaltered (Fig. 4b). Of note, only tumour-antigen-specific $CD4^+$ T cells devoid of XBP1 demonstrated maximal IFNγ production in the ovarian cancer microenvironment (Fig. 4c). XBP1-deficient intra-tumoral $CD8^+$ T cells also showed enhanced effector profiles, as evidenced by increased perforin and IFNγ expression compared with their wild-type counterparts (Extended Data Fig. 7h, i). Ablation of XBP1 did not influence glucose import, but did increase the mitochondrial membrane potential of $CD44^{high}CD4^+$ T cells at tumour sites, compared with the same T cell population from wild-type controls (Extended Data Fig. 7j, k). Expression of PD-1, CTLA-4 or TIM-3 was not altered (Extended Data Fig. 7l), which suggests that the enhanced effector profile of XBP1-deficient T cells is likely to be independent of common immune checkpoint signalling.

Overexpression of IFNγ emerged as a top biomarker of enhanced effector function by $CD4^+$ T cells that lack XBP1. We sought to validate these findings in the human context. Exposure to ascites derived from patients with ovarian cancer dampened IFNγ production by $CD4^+$ T cells (Extended Data Fig. 8a, b), but disabling IRE1α–XBP1 with 4μ8C partly alleviated this effect and enhanced T cell mitochondrial respiration (Extended Data Fig. 8c–e). 4μ8C had minimal effects on T cell viability or expression of T-bet or RORγt in $CD4^+$ T cells exposed to ascites (Extended Data Fig. 8f–h). Notably, enforcing SNAT1 expression in T cells also augmented their capacity to express IFNγ under glucose deprivation or exposure to ovarian cancer ascites (Extended Data Fig. 8i).

We next assessed whether IRE1α–XBP1 activation in T cells influenced malignant progression. The growth of ovarian tumours driven by p53 loss and mutant KRAS[27] was compromised in female mouse hosts that have a haematopoietic compartment devoid of XBP1 selectively in T cells (Fig. 4d). Reduced tumour growth was also evidenced in *Xbp1*$^{f/f}$*Cd4*^{cre} female mice challenged with ID8-based ovarian cancer cells in the flank (Extended Data Fig. 7m). Female mice that lack XBP1 in T cells showed decreased ascites accumulation and increased survival when challenged with metastatic ID8-*Defb29*/*Vegfa* tumours[28] (Fig. 4e). Similar results were observed when these tumours were developed in *Ern1*$^{f/f}$*Cd4*^{cre} mice (Extended Data Fig. 7n), which indicates that canonical IRE1α–XBP1 activation in T cells—rather than XBP1-independent IRE1α kinase signalling or regulated IRE1α-dependent mRNA decay[29]—is responsible for the delayed ovarian cancer progression observed in XBP1-deficient mouse hosts. *Ern1*$^{f/f}$*Cd4*^{cre} mice that bear parental ID8 tumours[30] that do not overexpress *Defb29* and *Vegfa* also showed reduced peritoneal carcinomatosis and extended survival, compared with their *Ern1*$^{f/f}$ littermate controls (Fig. 4f, g).

Here we present experimental evidence that indicates that ovarian cancer exploits the IRE1α–XBP1 arm of the UPR to cripple T cell metabolism and anti-tumour capacity. The UPR normally facilitates adaptation to ER stress under nutrient-rich conditions; our findings suggest that T cells experience maladaptive IRE1α–XBP1 activation within tumours in which glucose availability is restricted (Extended Data Fig. 9). We propose that disruption of ER homeostasis in intratumoral T cells operates as an integrated 'immunometabolic checkpoint' that influences adaptive immunity and malignant progression in cancer hosts.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0597-x.

Received: 8 April 2017; Accepted: 21 August 2018;
Published online 10 October 2018.

1. Chang, C. H. et al. Metabolic competition in the tumor microenvironment is a driver of cancer progression. *Cell* **162**, 1229–1241 (2015).
2. Ho, P. C. et al. Phosphoenolpyruvate is a metabolic checkpoint of anti-tumor T cell responses. *Cell* **162**, 1217–1228 (2015).
3. Scharping, N. E. et al. The tumor microenvironment represses T cell mitochondrial biogenesis to drive intratumoral T cell metabolic insufficiency and dysfunction. *Immunity* **45**, 374–388 (2016).
4. Anderson, K. G., Stromnes, I. M. & Greenberg, P. D. Obstacles posed by the tumor microenvironment to T cell activity: a case for synergistic therapies. *Cancer Cell* **31**, 311–325 (2017).
5. Chae, C. S., Teran-Cabanillas, E. & Cubillos-Ruiz, J. R. Dendritic cell rehab: new strategies to unleash therapeutic immunity in ovarian cancer. *Cancer Immunol. Immunother.* **66**, 969–977 (2017).
6. Matulonis, U. A. et al. Ovarian cancer. *Nat. Rev. Dis. Primers* **2**, 16061 (2016).
7. Hamanishi, J. et al. Safety and antitumor activity of anti-PD-1 antibody, nivolumab, in patients with platinum-resistant ovarian cancer. *J. Clin. Oncol.* **33**, 4015–4022 (2015).
8. Kershaw, M. H. et al. A phase I study on adoptive immunotherapy using gene-modified T cells for ovarian cancer. *Clin. Cancer Res.* **12**, 6106–6115 (2006).
9. Yoshida, H., Matsui, T., Yamamoto, A., Okada, T. & Mori, K. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell* **107**, 881–891 (2001).
10. Lee, A. H., Iwakoshi, N. N. & Glimcher, L. H. XBP-1 regulates a subset of endoplasmic reticulum resident chaperone genes in the unfolded protein response. *Mol. Cell. Biol.* **23**, 7448–7459 (2003).
11. Cubillos-Ruiz, J. R., Bettigole, S. E. & Glimcher, L. H. Tumorigenic and immunosuppressive effects of endoplasmic reticulum stress in cancer. *Cell* **168**, 692–706 (2017).
12. Cubillos-Ruiz, J. R. et al. ER stress sensor XBP1 controls anti-tumor immunity by disrupting dendritic cell homeostasis. *Cell* **161**, 1527–1538 (2015).
13. Yan, D., Wang, H. W., Bowman, R. L. & Joyce, J. A. STAT3 and STAT6 signaling pathways synergize to promote cathepsin secretion from macrophages via IRE1α activation. *Cell Rep.* **16**, 2914–2927 (2016).
14. Condamine, T. et al. Lectin-type oxidized LDL receptor-1 distinguishes population of human polymorphonuclear myeloid-derived suppressor cells in cancer patients. *Sci. Immunol.* **1**, aaf8943 (2016).
15. Kipps, E., Tan, D. S. & Kaye, S. B. Meeting the challenge of ascites in ovarian cancer: new avenues for therapy and research. *Nat. Rev. Cancer* **13**, 273–282 (2013).
16. Bamias, A. et al. Significant differences of lymphocytes isolated from ascites of patients with ovarian cancer compared to blood and tumor lymphocytes. Association of $CD3^+CD56^+$ cells with platinum resistance. *Gynecol. Oncol.* **106**, 75–81 (2007).
17. Curiel, T. J. et al. Specific recruitment of regulatory T cells in ovarian carcinoma fosters immune privilege and predicts reduced survival. *Nat. Med.* **10**, 942–949 (2004).
18. Lukesova, S. et al. Comparative study of various subpopulations of cytotoxic cells in blood and ascites from patients with ovarian carcinoma. *Contemp. Oncol. (Pozn)* **19**, 290–299 (2015).
19. Knutson, K. L. et al. Regulatory T cells, inherited variation, and clinical outcome in epithelial ovarian cancer. *Cancer Immunol. Immunother.* **64**, 1495–1504 (2015).
20. Kim, K. S. et al. Hypoxia enhances lysophosphatidic acid responsiveness in ovarian cancer cells and lysophosphatidic acid induces ovarian tumor metastasis in vivo. *Cancer Res.* **66**, 7983–7990 (2006).
21. Denzel, M. S. & Antebi, A. Hexosamine pathway and (ER) protein quality control. *Curr. Opin. Cell Biol.* **33**, 14–18 (2015).
22. Yang, C. et al. Glutamine oxidation maintains the TCA cycle and cell survival during impaired mitochondrial pyruvate transport. *Mol. Cell* **56**, 414–424 (2014).

23. Wellen, K. E. et al. The hexosamine biosynthetic pathway couples growth factor-induced glutamine uptake to glucose metabolism. *Genes Dev.* **24**, 2784–2799 (2010).
24. Wang, H. et al. Endoplasmic reticulum stress up-regulates Nedd4-2 to induce autophagy. *FASEB J.* **30**, 2549–2556 (2016).
25. Jeon, Y. J. et al. Regulation of glutamine carrier proteins by RNF5 determines breast cancer response to ER stress-inducing chemotherapies. *Cancer Cell* **27**, 354–369 (2015).
26. Hatanaka, T., Hatanaka, Y. & Setou, M. Regulation of amino acid transporter ATA2 by ubiquitin ligase Nedd4-2. *J. Biol. Chem.* **281**, 35922–35930 (2006).
27. Scarlett, U. K. et al. Ovarian cancer progression is controlled by phenotypic changes in dendritic cells. *J. Exp. Med.* **209**, 495–506 (2012).
28. Conejo-Garcia, J. R. et al. Tumor-infiltrating dendritic cell precursors recruited by a β-defensin contribute to vasculogenesis under the influence of Vegf-A. *Nat. Med.* **10**, 950–958 (2004).
29. So, J. S. et al. Silencing of lipid metabolism genes through IRE1α-mediated mRNA decay lowers plasma lipids in mice. *Cell Metab.* **16**, 487–499 (2012).
30. Roby, K. F. et al. Development of a syngeneic mouse model for events related to ovarian cancer. *Carcinogenesis* **21**, 585–591 (2000).

Acknowledgements Our research was supported by the Irvington Institute Fellowship Program of the Cancer Research Institute (J.R.C.-R.), the Ann Schreiber Mentored Investigator Award of the Ovarian Cancer Research Fund Alliance (J.R.C.-R.), the Ovarian Cancer Academy Early-Career Investigator Award W81XWH-16-1-0438 of the Department of Defense (J.R.C.-R.), the Stand Up to Cancer Innovative Research Grant SU2C-AACR-IRG-03-16 (J.R.C.-R.), the Jacquie Liggett Fellowship Award of Hearing the Ovarian Cancer Whisper (J.R.C.-R.), Weill Cornell Medicine Funds (J.R.C.-R. and L.H.G.), NIH grant R01CA112663 (L.H.G.) and NIH SIG grant 1S10 OD017992-01 (S.Z.) for the Orbitrap Fusion mass spectrometer. We thank T. Iwawaki at Kanazawa Medical University for sharing the *Ern1*-floxed mouse strain; J. McCormick for expert assistance with cell sorting; L. Cohen-Gould and J. Jimenez for electron microscopy analysis; G. Zhang, Z. Cheng and T. Su for metabolic tracing experiments; all members of the Weill Cornell Epigenomics Facility for assistance with RNA sequencing, T. Walther for help collecting patient samples; J. M. Pérez-Sáez and J. Trillo-Tinoco for assistance with some experimental analyses and helpful suggestions; and L. Cantley and M. Goncalves for sharing valuable instruments and resources. We also thank all members of the Cubillos-Ruiz, Morales and Glimcher laboratories for helpful suggestions and critical reading of this manuscript.

Reviewer information *Nature* thanks T. Curiel and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.S. designed and conducted most of the in vitro and in vivo experiments, analysed data and wrote the manuscript. T.A.S., C.-S.C., S.C., C.T., M.R., R.A.C., K.K.P., H.R.S., M.J.P.C. and J.P.C performed in vitro and in vivo experiments and analysed data. M.R.R performed in vivo experiments using p53/KRAS hosts and analysed data. C.K. and G.M. contributed to the design of certain Seahorse-related experiments and shared resources. S.E.B. performed mixed BM chimaeras, in vitro experiments and edited the manuscript. A.V.K. carried out computational analyses of RNA sequencing data. I.M. and S.Z. performed mass spectrometry experiments and analysed proteomics data. K.H. and D.Z. performed surgeries and provided patient specimens. P.C.R., G.A.R. and J.R.C.-G. contributed to the design of certain experiments, provided ideas and models, shared resources, analysed data and reviewed the manuscript. L.H.G. designed the research, provided models and resources, analysed data and reviewed the manuscript. J.R.C.-R. conceived the idea, designed and conducted the research, analysed data, wrote the manuscript and directed the project. J.R.C.-R. is lead senior author for this paper.

Competing interests J.R.C.-R. and L.H.G. are co-founders of and scientific advisors for Quentis Therapeutics. S.E.B. is co-founder and employee of Quentis Therapeutics. L.H.G. also serves on the board of directors of and holds equity in GlaxoSmithKline Pharmaceuticals.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0597-x.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0597-x.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.H.G. or J.R.C.-R.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment, unless stated otherwise.

Patient-derived specimens. Stage III and stage IV human ovarian carcinoma tumours and malignant ascites fluid were procured through Surgical Pathology at Weill Cornell Medicine and Memorial Sloan Kettering Cancer Center. The Weill Cornell Medicine IRB conducted a review of the project described and determined that the activities do not constitute human subjects research as these specimens were classified as surgical discard and remained completely de-identified. Collection and analysis of ovarian cancer specimens at Memorial Sloan Kettering Cancer Center were approved by an IRB protocol and were obtained under informed consent. Tumour single-cell suspensions were prepared as previously described[28]. Malignant ascites samples were centrifuged for 10 min at 1,300 r.p.m. Ascites supernatants were collected, depleted of cells by passing through 0.22-μm filters and stored frozen at −80 °C as small aliquots until use. Red blood cells in cell pellets were lysed with ACK lysing buffer (Gibco). Tumour infiltrating T cells ($CD45^+CD20^-CD14^-CD3^+CD4^+$ or $CD45^+CD20^-CD14^-CD3^+CD8^+$) were sorted from tumour single-cell suspensions or malignant ascites using a BD FACS Aria II SORP cell sorter at Flow Cytometry Core Facility in Weill Cornell Medicine. Dead cells were excluded using the LIVE/DEAD Fixable Yellow Dead Cell Stain Kit (Life Technologies). All specimens of ovarian cancer used and analysed in this study are described in Supplementary Table 1.

Transgenic mice and experimental ovarian cancer models. Mice devoid of XBP1 or IRE1α selectively in T cells were generated by breeding *Xbp1*$^{f/f}$ or *Ern1*$^{f/f}$ transgenic mice with the *Cd4*^{cre} strain[31,32]. All mouse strains are on a full C57BL/6 background and female mice were used at 6–8 weeks of age. Mice were housed in specific pathogen-free animal facilities at Weill Cornell Medical College and Memorial Sloan Kettering Cancer Center. Mice were handled in compliance with Weill Cornell Institutional Animal Care and Use Committees procedures and guidelines. Gene-deficient mice were housed separately from wild-type littermate controls after genotyping at 3–4 weeks of age. Functional and survival experiments were conducted using age-matched littermate controls. In vivo experiments included three to six mice per group, based on transgenic genotype and sex availability. No randomization or blinding method was used for animal studies. Wild-type C57BL/6 mice were purchased from Jackson Laboratories.

'p53/KRAS' double transgenic animals were generated by crossing loxp-STOP-loxp-*Kras*$^{G12D/+}$ mice with *p53*$^{loxp/loxp}$ mice as previously reported[27]. p53/KRAS mice were irradiated two consecutive days with 6.5 Gy followed by bone marrow transplantation from either *Xbp1*$^{f/f}$ or *Xbp1*$^{f/f}$*Cd4*^{cre} mice. At eight weeks after bone marrow reconstitution, ovarian tumours were initiated by injecting Cre recombinase-expressing adenovirus into the ovarian bursa. Tumour size was measured and volume was estimated using the formula $V = 0.5$ (length × width^2). Mice were euthanized nine weeks after tumour initiation and tumours were resected. Specimens were minced, digested for 1 h at 37 °C in RPMI containing 2 mg/ml collagenase D and 1 mg/ml DNase I and then passed through a 70-μm strainer to obtain single-cell suspensions. Red blood cells were eliminated using ACK lysis buffer (Gibco), and cell suspensions of $5–10 \times 10^7$ cells/ml were stored in freezing medium (FBS containing 10% DMSO) at −80 °C.

Parental ID8 cells expressing luciferase (ID8-Luc) or ovalbumin protein (ID8-OVA), and aggressive ID8-*Defb29/Vegfa* cancer cell-lines were generated as previously described[28,30]. No authentication method was performed for the mouse cell lines used. All cell lines routinely tested negative for mycoplasma contamination using the Mycoplasma Detection Kit (Lonza) and maintained in medium containing plasmocin (Invivogen) for prophylactic purpose. In brief, 1.5×10^6 tumour cells were injected into the peritoneal cavity or right flank using Matrigel (Corning) of wild-type or transgenic mice described above. For live bioluminescent imaging, mice were given a single intraperitoneal injection of VivoGlo luciferin (Promega) and imaged with an IVIS Spectrum In Vivo imaging system at the Weill Cornell Research Animal Resource Center. To assess antigen-specific anti-tumour responses, *Xbp1*$^{f/f}$ or *Xbp1*$^{f/f}$*Cd4*^{cre} female mice were independently challenged with ID8-Luc or ID8-OVA. Tumour-associated $CD4^+$ T cells were simultaneously isolated from the peritoneal cavity of all mice 35 days after tumour challenge. Then, 5×10^5 T cells were co-cultured for 3 days with 1×10^5 BMDCs previously pulsed with $OVA_{323–339}$ peptide (1 μg/ml, InvivoGen). Cell culture supernatants were collected and IFNγ concentration was determined by enzyme-linked immunosorbent assay (ELISA) using the Ready-SET-Go kit (eBioscience).

To generate bone marrow chimaeras, recipient $CD45.1^+$ C57BL/6 mice were fed with a Sulfatrim diet during 7 days before irradiation. Animals were then exposed to a single lethal dose of irradiation (10 Gy). Twenty-four hours later, 5×10^6 total bone marrow cells from a 1:1 mixture of wild-type $CD45.1^+$ C57BL/6 and $CD45.2^+$ *Xbp1*$^{f/f}$ or *Xbp1*$^{f/f}$*Vav1*^{cre} bone marrow cells were injected intravenously into the irradiated recipient hosts. Mice were maintained on a Sulfatrim-based diet until the donor-derived cells in bone marrow and spleen were analysed by flow cytometry.

$CD4^+$ T cell isolation and in vitro cell culture. Peripheral blood mononuclear cells from cancer-free female donors (New York Blood Center) were isolated by density-gradient centrifugation using Ficoll (GE Healthcare). $CD4^+$ T cells were then enriched by negative selection using the human $CD4^+$ T cell isolation kit (Miltenyi Biotech). Unless noted otherwise, $CD4^+$ T cells were stimulated with Dynabeads human T-activator CD3 and CD28 (Thermo Fisher Scientific) at a 1:1 ratio for 16 h in the presence or absence of supernatants from ovarian cancer ascites, at the indicated concentrations. Depending on sample availability, all functional assays involving glucose uptake, IRE1α–XBP1 activation, bioenergetics and INFγ secretion were performed using $CD4^+$ T cells obtained from at least 3 distinct donors that were independently exposed to ascites supernatants obtained from 2–7 different patients with ovarian cancer.

Mouse tumour-infiltrating T cells ($CD45^+CD3^+CD4^+$ or $CD45^+CD3^+CD8^+$) were sorted from various sources, including single-cell suspension of ovarian tumours driven by p53 and KRAS, spleen, draining lymph nodes and peritoneal wash, malignant ascites and spleen from mice bearing aggressive ID8-*Defb29/Vegfa* ovarian cancer. Mouse splenic naive $CD4^+$ T cells were isolated by negative selection (Miltenyi Biotech) and then activated with plate-bound anti-CD3ε (145-2C11, 5 μg/ml) and soluble anti-CD28 (37.51, 1 μg/ml; BD Pharmingen) antibodies for 48 h. All $CD4^+$ T cells were cultured in complete RPMI, which is glucose-rich RPMI-1640 medium (Corning) further supplemented with 10% heat-inactivated FBS (Atlanta Biologicals), 2 mM L-glutamine (Corning), 25 mM HEPES, pH 7.2–7.6 (Corning), non-essential amino acids (Corning), 1 mM sodium pyruvate (Gibco), 100 IU penicillin and 100 μg/ml streptomycin (Corning) and 55 μM 2-mercaptoethanol (Gibco).

RNA extraction, reverse-transcription PCR (RT–PCR), qPCR and XBP1 splicing assays. Total RNA was isolated using RNeasy Mini kit or QIAzol lysis reagent (Qiagen) according to the manufacturer's instructions. Between 0.1 and 1 μg of RNA was reverse-transcribed to generate cDNA using the qScript cDNA synthesis kit (Quantabio). qPCR was performed using PerfeCTa SYBR green fastmix (Quantabio) and TaqMan Universal PCR master mix (Life Technologies) on a QuantStudio 6 Flex real-time PCR system (Applied Biosystems). Normalized gene expression was calculated by comparative threshold cycle method using *ACTB, GAPDH* or *Actb* as a control.

The *XBP1* splicing assay was performed as previously described[33]. The PCR products were separated by electrophoresis through a 2.5% agarose gel and visualized by ethidium bromide staining. The ImageJ software for densitometry was used for calculating the frequency of *XBP1* mRNA splicing (%) as follows: intensity of *XBP1s* band divided by (intensity of *XBP1s* + *XBP1u* band) × 100.

All primers and Taqman probes used in this study are described in Supplementary Table 3.

Analyses based on flow cytometry. Flow cytometry was conducted using fluorochrome-conjugated antibodies purchased from BioLegend, unless stated otherwise. For staining of mouse cells we used: anti-CD45 (30-F11), anti-CD4 (RM4-5), anti-CD8 (53-6.7), anti-CD3 (145-2C11), anti-TCRβ (H57-597), anti-CD44 (IM7), anti-CD62L (MEL-14), anti-CD11b (M1/70), anti-CD11c (N418), anti-CD19 (6D5), anti-F4/80 (BM8), anti-Ly6C (HK1.4), anti-Ly6G (1A8), anti-I-A/I-E (M5/114.15.2), anti-IFNγ (XMG1.2), anti-IL-17A (TC11-18H10.1), anti-FoxP3 (150D), anti-PD-1 (29F.1A12), anti-CTLA4 (UC10-4B9), anti-TIM3 (B8.2C12), TruStain fcX CD16/32 (93); anti-perforin (17-9392-80, eBioscience). To stain human cells we used: anti-CD45 (HI30), anti-CD3 (HIT3), anti-CD4 (A161A1), anti-CD8 (HIT8a), anti-CD20 (2H7), anti-CD14 (63D3), anti-IFNγ (4S.B3), anti-T-bet (4B10), TruStain FcX solution (422302); anti-XBP1s (Q3-695; BD Pharmingen); anti-RORγt (AFKJS-9, eBioscience). Staining of transcription factors XBP1s and FoxP3 and cytokines was carried out using FoxP3/transcription factor staining buffer set (eBioscience) according to the manufacturer's instructions. For in vitro glucose uptake experiments, 2×10^5 human $CD4^+$ T cells were incubated with 20 μM 2-NBDG (Thermo Fisher Scientific) in glucose-free medium for 30 min at 37 °C. To measure levels of intracellular reactive oxygen species, cells were stained with 20 μM DCFDA (Abcam) for 30 min at 37 °C. The surface expression of GLUT1 in T cells residing in human ovarian cancer ascites was analysed using GLUT1.RBD.GFP (Metafora Biosystems) according to the manufacturer's instructions. In brief, 2×10^5 cells were labelled with 5 μl GLUT1.RBD.GFP in 100 μl of complete RPMI containing 0.09% NaN_3 and 1 mM EDTA and incubated for 30 min at 37 °C. Then cells were washed with PBS followed by cell-surface staining. To characterize ovarian cancer-infiltrating $CD4^+$ T cells, peritoneal wash samples from *Xbp1*$^{f/f}$ or *Xbp1*$^{f/f}$*Cd4*^{cre} female mice bearing metastatic ID8-*Defb29/Vegfa* ovarian cancer for 20–23 days were used. For intracellular cytokine staining, 5×10^6 cells from peritoneal wash samples were stimulated for 6 h in complete RPMI containing cell activation cocktail with brefeldin A (Biolegend). For 2-NBDG uptake in vivo, tumour-bearing mice were injected intraperitoneally with 100 μg 2-NBDG per mouse diluted in PBS and then, mice were euthanized 30 min after injection for collecting peritoneal cells. To measure mitochondrial membrane potential, peritoneal wash cells were stained with 200 nM TMRE (Thermo Fisher Scientific)

dye for 30 min at 37 °C followed by cell surface staining. To visualize proliferation assays, T cells were labelled with 5 μM CellTrace Violet (Thermo Fisher Scientific) at 37 °C for 30 min. The cell cycle of preactivated $CD4^+$ T cells was analysed by staining DNA with 50 μg/ml propidium iodide solution containing 0.5 μg/ml RNase I after fixing the cells in cold 70% ethanol for 1 h at 4 °C. Forty-eight hours post-activation, $CD4^+$ T cells were loaded with MitoTracker (20 nM, Thermo Fisher Scientific) to measure mitochondrial mass. All events were acquired on an LSRII (BD Biosciences) instrument and data were analysed with FlowJo software (TreeStar).

Immunoblot analysis. T cells were washed twice in 1× cold PBS and cell pellets were lysed using RIPA lysis and extraction buffer (Thermo Fisher Scientific) supplemented with a protease and phosphatase inhibitor tablet (Roche). Homogenates were centrifuged at 14,000 r.p.m. for 30 min at 4 °C, and the supernatants were collected. Protein concentrations were determined using a BCA protein assay kit (Thermo Fisher Scientific). Equivalent amounts of protein were separated by SDS–PAGE and transferred onto PVDF membranes following the standard protocol. The following antibodies were used: anti-GLUT1 (D3J3A), anti-ACTB (4967S), anti-ASCT2 (V501) and anti-SNAT1/SLC38A1 (D9L2P) (Cell Signaling Technologies); anti-PGC-1α (H-300) and anti-SNAT2 (C-6) (Santa Cruz Biotechnology); and goat anti-rabbit and mouse secondary antibodies conjugated with HRP (Thermo Fisher Scientific). SuperSignal West Pico and Femto chemiluminescent substrates (Thermo Fisher Scientific) were used to image blots in a FluorChemE instrument (ProteinSimple). The proteasomal inhibitor MG132 (10 μM, Sigma) and translation inhibitor cycloheximide (CHX, 50 μg/ml, Sigma) were used to assess protein abundance and stability. Densitometric quantification was performed using Image J software (NIH).

Seahorse analyses. Purified $CD4^+$ T cells from cancer-free female donors were stimulated with human T-activator CD3 and CD28 Dynabeads (Life Technologies) in complete RPMI medium containing 10% FBS, in the presence or absence of supernatants from ovarian cancer ascites for 16 h. The effect of IRE1α inhibition using 4μ8C (10 μM, EMD Millipore) or GlcNAc supplementation (10 mM, Sigma) was examined by pre-treating T cells 1–2 h before addition of human ovarian cancer ascites. DMKG (5 mM, Sigma) was added in the cell culture during the final 4 h of incubation. After incubation, T cells were collected, thoroughly washed and then subjected to Seahorse analyses using non-buffered XF base medium containing 25 mM glucose (Sigma), 2 mM L-glutamine, and 1 mM sodium pyruvate but lacking serum or ascites. XF96 cell-culture microplates were coated with CellTak (Corning) before analysis according to the manufacturer's instructions and washed twice with distilled water. Unless stated otherwise, 1.5×10^5 T cells were plated and OCR and ECAR measurements were analysed on an XFe96 Extracellular Flux Analyzer (Agilent). After basal OCR and ECAR measurements were obtained, an OCR trace was recorded in response to oligomycin (1 μM), carbonyl cyanide-*p*-(trifluoromethoxy) phenylhydrazone (FCCP, 1 μM), and rotenone and antimycin (0.5 μM each) following the XF Cell Mito Stress test kit (Agilent). To evaluate mitochondrial fuel usage, T cells were seeded in the Seahorse medium containing 2 mM L-glutamine ± 25 mM glucose. After recording basal OCR and ECAR measurements, cells were injected with corresponding base medium (control), UK5099 (2 μM), BPTES (3 μM) and etomoxir (4 μM) using the XF Mito Fuel Flex test kit (Agilent) followed by oligomycin, FCCP, and rotenone and antimycin injection. Metabolic parameters were calculated as follow: basal OCR = last rate measurement before oligomycin injection – minimum rate measurement after rotenone and antimycin injection; maximal OCR = maximum rate measurement after FCCP injection – minimum rate measurement after rotenone and antimycin injection. Typically, 3–6 technical replicates per each sample were examined. After analysis, the cell numbers of each well were determined by nuclear DNA staining with Hoechst 33342 (Sigma) and OCR and ECAR values were normalized accordingly.

Glutamine tracing experiments. For stable isotope tracer experiments, naive splenic $CD4^+$ T cells isolated from wild-type or XBP1-deficient mice were activated by CD3 and CD28 stimulation for 48 h, followed by culture in the presence or absence of glucose for 4.5 h, and then pulsed with [U-^{13}C]glutamine (Sigma) for an additional 1.5 h in the same culture condition. After washing twice with cold PBS, metabolites were extracted from the cells by methanol extraction method[34]. In brief, pre-cooled 80% methanol (1 ml) was added to each sample and kept in −80 °C for 20 min. Samples were then centrifuged at 4 °C for 5 min at 14,000 r.p.m. The supernatants were extracted and normalized based on cell amount. Targeted liquid chromatography–mass spectrometry (LC–MS) analyses were performed on a Q Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific) coupled to a Vanquish UPLC system (Thermo Fisher Scientific). The Q Exactive operated in polarity-switching mode. A Sequant ZIC-HILIC column (2.1 mm i.d. × 150 mm, Merck) was used for separation of metabolites. Flow rate was set at 150 μl/min. Buffers consisted of 100% acetonitrile for mobile A, and 0.1% NH_4OH and 20 mM CH_3COONH_4 in water for mobile B. Gradient ran from 85% to 30% A in 20 min followed by a wash with 30% A and re-equilibration at 85% A. Metabolites were identified on the basis of exact mass within 5 p.p.m. and standard retention times. Relative metabolite quantitation was performed based on peak area for each metabolite. Tracing experiments were performed at the Proteomics and Metabolomics Core Facility of Weill Cornell Medicine.

Immunofluorescence and confocal microscopy. Cells were transferred onto poly-L-lysine pre-coated glass coverslips (Neuvitro) and incubated at 37 °C for 30 min. The coverslips were washed with cold PBS three times between each step. Cells were immediately fixed with ice-cold acetone for 10 min at room temperature. Then, the coverslips were blocked for 1 h in PBS containing 0.25% Triton X-100 and 5% FBS at 4 °C, followed by incubation at 4 °C overnight with primary antibodies rabbit anti-ASCT2 (V501, Cell Signaling Technologies, 1:200) and mouse anti-SNAT2 (C-6) (Santa Cruz Biotechnology, 1:200) in PBS containing 5% FBS, protected from light. Then, secondary antibodies Alexa Fluor 488-conjugated goat anti-rabbit IgG and Alexa Fluor 594-conjugated goat anti-mouse IgG (Molecular Probes, 1:400) were added for 1 h at room temperature in the dark. Cells were counterstained with 4′,6-diamidino-2-phenylindole (DAPI, Thermo Fisher Scientific, 0.5 μg/ml) for 5 min at room temperature in the dark. After washing and removing excess solution, the flipped coverslips were placed on the mounting medium (Southern Biotechnology). Slides were allowed to dry in the dark for 1 h in a humid chamber at room temperature. Slides were then sealed with fingernail polish before examination. Digital confocal images were captured on a Zeiss LSM 880 Confocal Microscope with the Airy Scan high-resolution detector at the Weill Cornell CLC Imaging Core Facility. Image J software was used to determine the average pixel intensity of 16-bit greyscale images from each channel, individually. A region of interest was drawn around an individual cell that was not in contact with other cells, and the mean intensity was recorded. Approximately 50 cells in a field from three independent slides were quantified to calculate average fluorescence intensity ($n = 150$ total cells).

Retroviral transduction. The coding region of *SLC38A1* (NM_001278387.1), the gene encoding human SNAT1, was amplified using cDNA that originated from healthy peripheral blood mononuclear cells as a template and the following PCR primers: 5′-GAATTCGCCACCATGATGCATTTCAAAAGTGGACTCGA-3′ (forward); 5′-GCGGCCGCTCAGTGGCCTTCGTCACTACTCG-3′ (reverse). The PCR product was cloned into the pBMN-I-GFP retroviral-expressing vector (Addgene) and subsequently transfected into the Phoenix-AMPHO (ATCC) retrovirus producer cell line using Lipofectamine 3000 (Invitrogen) to generate retroviruses. Cell-culture supernatants containing virus were collected at 48 h and 72 h post-transfection. The virus was then bound to 24-well non-tissue-culture-treated plates previously coated with RetroNectin (70 μg/ml, Takara Bio) by centrifugation at 2,000*g* for 2 h at 37 °C. Human $CD4^+$ T cells were stimulated with anti-CD3 and anti-CD28 Dynabeads (Life Technologies) in the presence of recombinant human IL-2 (50 units per ml, PeproTech) for 36 h before viral transduction. Typically, 1×10^6 T cells were added to each virus-coated well by brief centrifugation at a final concentration of 2×10^6 cells per ml containing 50 units per ml of IL-2 and second transduction was followed 24 h later as the same strategy. GFP^+ cells were sorted 48 h post-transfection and expanded with anti-CD3 and anti-CD28 stimulation in the presence of IL-2 for additional 48 h, for downstream assays.

ELISA. Eighty thousand human $CD4^+$ T cells were activated for 12 h and subsequently incubated for additional 36 h in the presence of 25% ascites. To test the effect of IRE1α inhibition on IFNγ secretion, T cells were activated for 36 h, treated with 4μ8C at 10 μM (or DMSO control) 2 h before adding 25% human ovarian cancer ascites, and incubated for additional 10 h. The concentration of IFNγ in culture supernatants was determined by ELISA using the Ready-SET-Go kit (eBioscience). IFNγ was undetectable in ascites supernatants when used at 25%. Cell viability and counts were comparable in all cases.

RNA sequencing and transcriptional profiling. Tumour infiltrating $CD4^+$ T cells ($CD45^+CD3^+CD4^+$) were sorted from peritoneal wash samples of *Xbp1*$^{f/f}$ or *Xbp1*$^{f/f}$*Cd4*^{cre} female mice bearing aggressive ID8-*Defb29*/*Vegfa* ovarian cancer for 20 days. $CD44^{high}CD62L^{low}CD4^+$ T cells were sorted from the spleen of naive *Xbp1*$^{f/f}$ or *Xbp1*$^{f/f}$*Cd4*^{cre} female mice. miRVana miRNA isolation kit (Ambion) and RNeasy MinElute kit (Qiagen) were used for total RNA isolation and concentration. All samples were passed RNA quality controls examined by Agilent Bioanalyzer 2100, and mRNA libraries were generated and sequenced at the Weill Cornell Epigenomics Core Facility.

Reads produced from 51-bp single-end sequencing runs were aligned against mouse genome (mm9) using Bowtie v.0.12.8 algorithm[35]. Mouse mm9 transcriptome information was obtained from UCSC Genome Browser[36] and the RSEM algorithm[37] was used to calculate number of aligned tags for each gene. Differential expression between two groups were tested by EdgeR[38] and genes that had at least 50 raw counts and passed FDR cut-off of 15% with at least 1.5-fold difference were considered significant. Official symbol and gene description information was obtained from NCBI Entrez information[39]. Normalized expression values of reads per kilobase of transcript per million mapped reads were generated by EdgeR and used to demonstrate gene expression across samples as a colour-coded fold change in expression in a sample, versus average expression across all samples. Functional enrichment analysis was done using Ingenuity Pathway Analysis (Qiagen).

Transmission electron microscopy. Pre-activated mouse CD4$^+$ T cells were fixed, dehydrated, embedded and sectioned for electron microscopy analysis following a standard protocol at Weill Cornell CLC Imaging Core Facility. Sections were viewed on a JEM 1400 Transmission Electron Microscope (JEOL) operated at 120 kV and digital images were acquired with a Veleta 2K × 2K charge-coupled device camera (Olympus-SIS).

Protein glycosylation assessment using liquid chromatography–tandem mass spectrometry. LC–MS-grade formic acid, acetonitrile (ACN) and water were purchased from Fisher Chemical. Trifluoroacetic acid (TFA) was acquired from Fluka. Dithiothreitol and iodoacetomide were purchased from Roche and Acros Organics, respectively.

Protein extraction, digestion and glyco-peptide enrichment. Samples were individually prepared for future peptide and protein identification and comparison. Human CD4$^+$ T cells were activated with anti-CD3 and anti-CD28 Dynabeads ± 50% ascites supernatants for 16 h. Cells were lysed using RIPA buffer and protein quantification was performed using the Pierce BCA Protein Assay kit (Thermo Fisher Scientific). Cell lysates (0.8 mg) were enriched for glycoproteins using the Con A-based Pierce Glycoprotein Isolation kit (Thermo Fisher Scientific). Enriched fractions were then concentrated by centrifugation and buffer exchange using the Amicon Ultra-0.5 Centrifugal Filter Unit with Ultracel-3 membrane (Millipore), according to the manufacturer's protocol, with the exchange buffer consisting of 4 M urea, 1 M thiourea and 50 mM TEAB at pH 8.5. Proteins were reduced with 10 mM dithiothreitol, incubated at 34 °C for 1 h, then alkylated with 58 mM iodoacetamide for 45 min in the dark at room temperature and then quenched by a final addition of 36 mM dithiothreitol. The solutions were then diluted with 50mM ammonium bicarbonate (pH 8.0) to a final buffer concentration of 1M urea before trypsin digestion. Each sample was digested with 0.8 μg of trypsin for 18 h at 37 °C. The digestion was stopped by addition of TFA to a final pH 2.2–2.5. The samples were then desalted with SOLA HRP SPE Cartridge (Thermo Fisher Scientific). First, the cartridges were conditioned with 1 × 0.5 ml 90% methanol, 0.1% TFA and equilibrated with 2 × 0.5 ml 0.1% TFA. The samples were diluted 1:1 with 0.2% TFA and were run slowly through cartridges. After washing with 2 × 0.5 ml of equilibration solution, peptides were eluted by 1 × 0.5 ml of 50% ACN, 0.1% TFA and dried in a speed-vacuum centrifuge. Glycosylated peptides were enriched on SOLA SAX SPE Cartridge (Thermo Fisher Scientific). Cartridges were conditioned with 1 × 1.0 ml of 100% ACN followed by 3 × 1.0 ml of 0.1 M triethylammonium acetate buffer (pH 7.0), washed with 1 × 1.0 ml ddH_2O and equilibrated with 1 × 1.0 ml 95% ACN, 1.0% TFA. Samples were reconstituted in 60 μl of 50% ACN, 0.1% TFA and loaded onto columns right after the equilibration step allowing slow flow-through. Cartridges were washed three times with 1.0 ml of equilibration solution and peptides were eluted twice with 0.6 ml of 50% ACN, 0.1% TFA, after which they were dried down in a speed-vacuum centrifuge for further use.

Nano-scale reverse-phase chromatography and tandem mass spectrometry. The nano-LC–MS/MS analysis was carried out using UltiMate3000 RSLCnano (Dionex) coupled to an Orbitrap Fusion (Thermo-Fisher Scientific) mass spectrometer equipped with a nanospray Flex Ion Source. Each sample was reconstituted in 22 μl of 0.5% formic acid and 10 μl was loaded onto a Acclaim PepMap 100 C18 trap column (5 μm, 100 μm × 20 mm, 100 Å, Thermo Fisher Scientific) with nanoViper Fittings at 20 μl/min of 0.5% formic acid for on-line desalting. After 2 min, the valve switched to allow peptides to be separated on an Acclaim PepMap C18 nano column (3 μm, 75 μm × 25 cm, Thermo Fisher Scientific). Mobile phase A consisted of 2% ACN, 0.1% formic acid in water, mobile phase B was 95% ACN, 0.1% formic acid in water and the 120 min gradient was as follows: 5% to 23% to 35% B at 300 nl/min (3 to 83 to 123 min, respectively), followed by a 9-min ramping to 90% B, a 9-min hold at 90% B and quick switch to 7% B in 1 min. The column was re-equilibrated with 5% B for 20 min before the next run. A 10-fmol injection of standard BSA digest mixture with a short 30-min gradient was run between every sample for quality-control purposes.

The Orbitrap Fusion instrument was operating in positive-ion mode with nanospray voltage set at 1.7 kV and source temperature at 275 °C. External calibration for Fourier transform, ion trap and quadrupole mass analysers was performed before the analysis. The Orbitrap full MS survey scan (m/z 400–1,800) was followed by top-3-s, data-dependent higher collision dissociation (HCD) product-dependent electron-transfer dissociation (ETD) MS/MS scans for precursor peptides with 2–8 charges above a threshold-ion count of 50,000 with normalized collision energy of 32%. Mass spectrometry survey scans were acquired at a resolving power of 120,000 (full-width at half maximum at m/z 200), with automatic gain control (AGC) = 2×10^5 and maximum injection time (maximum IT) = 50 ms, and HCD MS/MS scans at resolution 30,000 with AGC = 5×10^4, maximum IT = 60 ms and with Q isolation window (m/z) at 3 for the mass range m/z 105–2,000. Dynamic exclusion parameters were set at 1 within 60-s exclusion duration with ± 10 p.p.m. exclusion-mass width. The product-ion trigger list consisted of peaks at 204.0867 Da (HexNAc oxonium ion), 138.0545 Da (HexNAc fragment), 366.1396 Da (Hex-HexNAc oxonium ions) and 274.0927 Da (dehydrated *N*-acetylneuraminic acid). If one of the HCD product ions in the list was detected, two charge-dependent ETD MS/MS scans with HCD supplementary activation (SA for electron transfer and higher-energy collision dissociation (EThcD) scan) on the same precursor ion were triggered and collected in a linear ion trap. For doubly charged precursors, the ETD reaction time was set at 150 ms and the SA energy was set at 25%, and the same parameters set at 125 ms and 20%, respectively, were used for higher-charged precursors. For both ion-triggered scans, the fluoranthene ETD reagent target was set at 2×10^5, the AGC target at 1×10^4, maximum IT at 105 ms and isolation window at 3. All data were acquired under Xcalibur 3.0 operation software and Orbitrap Fusion Tune Application v.2.1 (Thermo Fisher Scientific).

Data processing, protein identification and analysis. All mass spectrometry and MS/MS raw spectra from each sample were searched using Byonics v.2.8.2 (Protein Metrics) using *Homo sapiens* protein database containing 133,840 sequences and downloaded from Uniprot TrEMLB on 4 January 2016. The peptide search parameters were as follows: two missed cleavage for full trypsin digestion with fixed carbamidomethyl modification of cysteine, variable modifications of methionine oxidation and deamidation on asparagine and glutamine residues. The peptide-mass tolerance was 10 p.p.m. and fragment-mass tolerance values for HCD and EThcD spectra were 0.05 Da and 0.6 Da, respectively. The maximum number of common and rare modifications were both set at two. The glycan search was performed against a list of 309 mammalian *N*-linked glycans. Identified peptides were filtered for maximum 2% FDR or 50 hits to the reverse database.

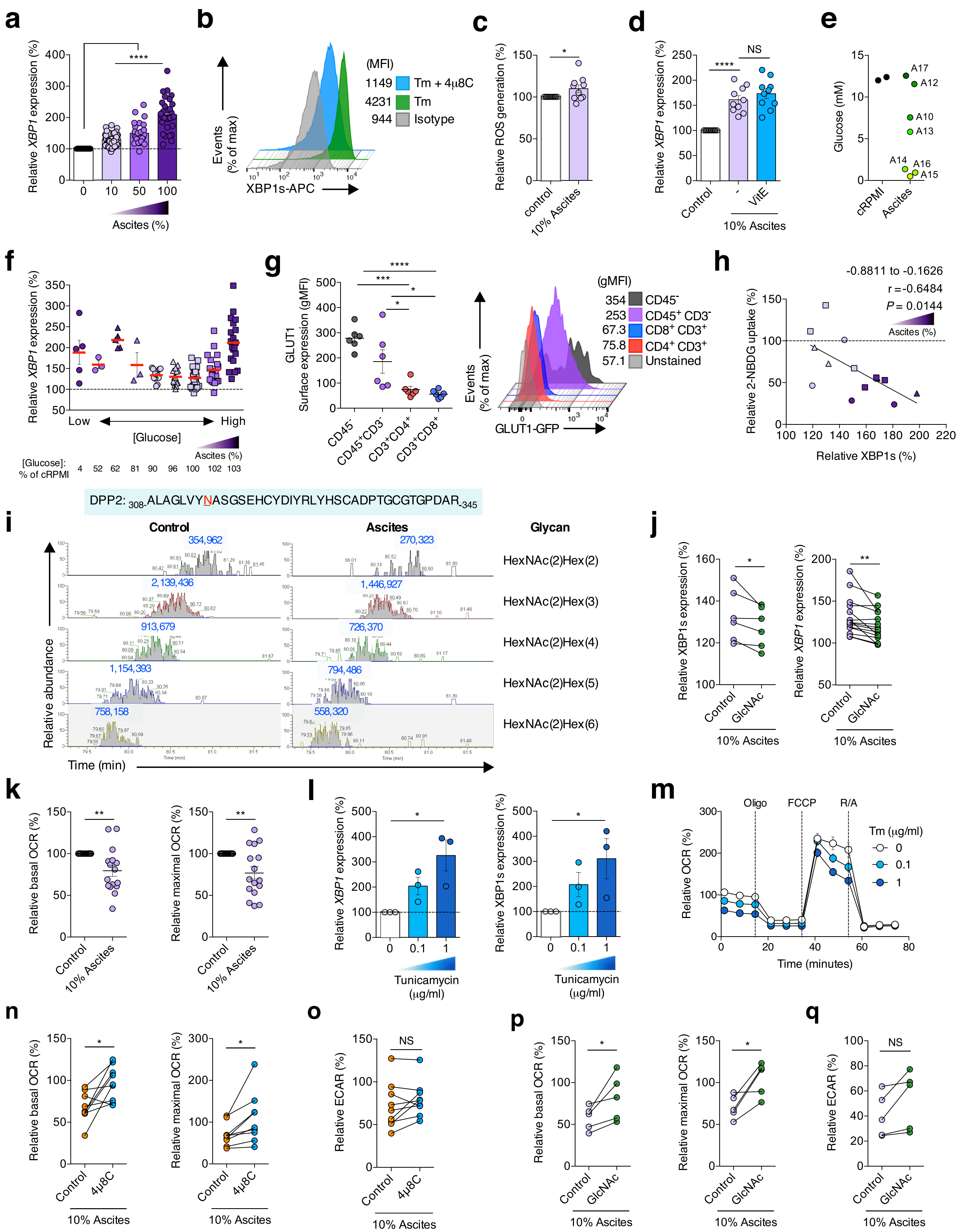
Statistical analyses. All statistical analyses were performed using GraphPad Prism 6 software. Significance for pairwise correlation analysis was calculated using the Spearman's correlation coefficient (r). Comparisons between two groups were assessed using unpaired or paired (for matched comparisons) two-tailed Student's t-test, or non-parametric Mann–Whitney U-test. Multiple comparisons were assessed by one-way ANOVA, including Tukey's or Bonferroni's multiple comparisons tests. Survival rates were compared using the log-rank test. Data are presented as mean ± s.e.m. P values of <0.05 were considered to be statistically significant.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

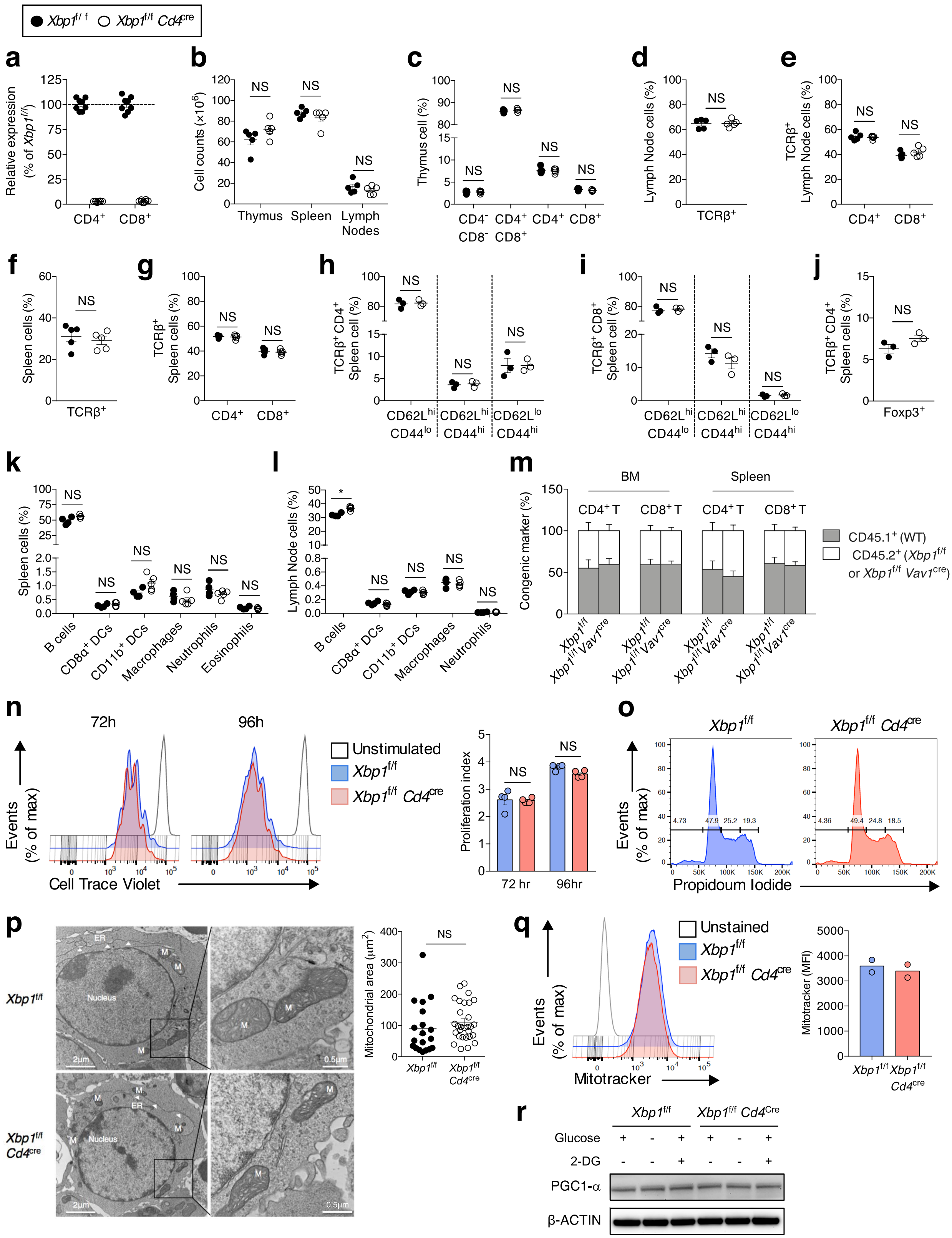
Source Data are provided for Figs. 1–4 and Extended Data Figs. 1–8. The NCBI GEO (Gene Expression Omnibus) accession number for RNA sequencing data reported in this paper is GSE118430. The datasets generated during the current study are available from the corresponding authors upon reasonable request.

31. Lee, A. H., Scapa, E. F., Cohen, D. E. & Glimcher, L. H. Regulation of hepatic lipogenesis by the transcription factor XBP1. *Science* **320**, 1492–1496 (2008).
32. Iwawaki, T., Akai, R., Yamanaka, S. & Kohno, K. Function of IRE1 alpha in the placenta is essential for placental development and embryonic viability. *Proc. Natl Acad. Sci. USA* **106**, 16657–16662 (2009).
33. Lee, A. H., Iwakoshi, N. N., Anderson, K. C. & Glimcher, L. H. Proteasome inhibitors disrupt the unfolded protein response in myeloma cells. *Proc. Natl Acad. Sci. USA* **100**, 9946–9951 (2003).
34. Yuan, M., Breitkopf, S. B., Yang, X. & Asara, J. M. A positive/negative ion-switching, targeted mass spectrometry-based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue. *Nat. Protoc.* **7**, 872–881 (2012).
35. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
36. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
37. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
38. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
39. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **39**, D52–D57 (2011).



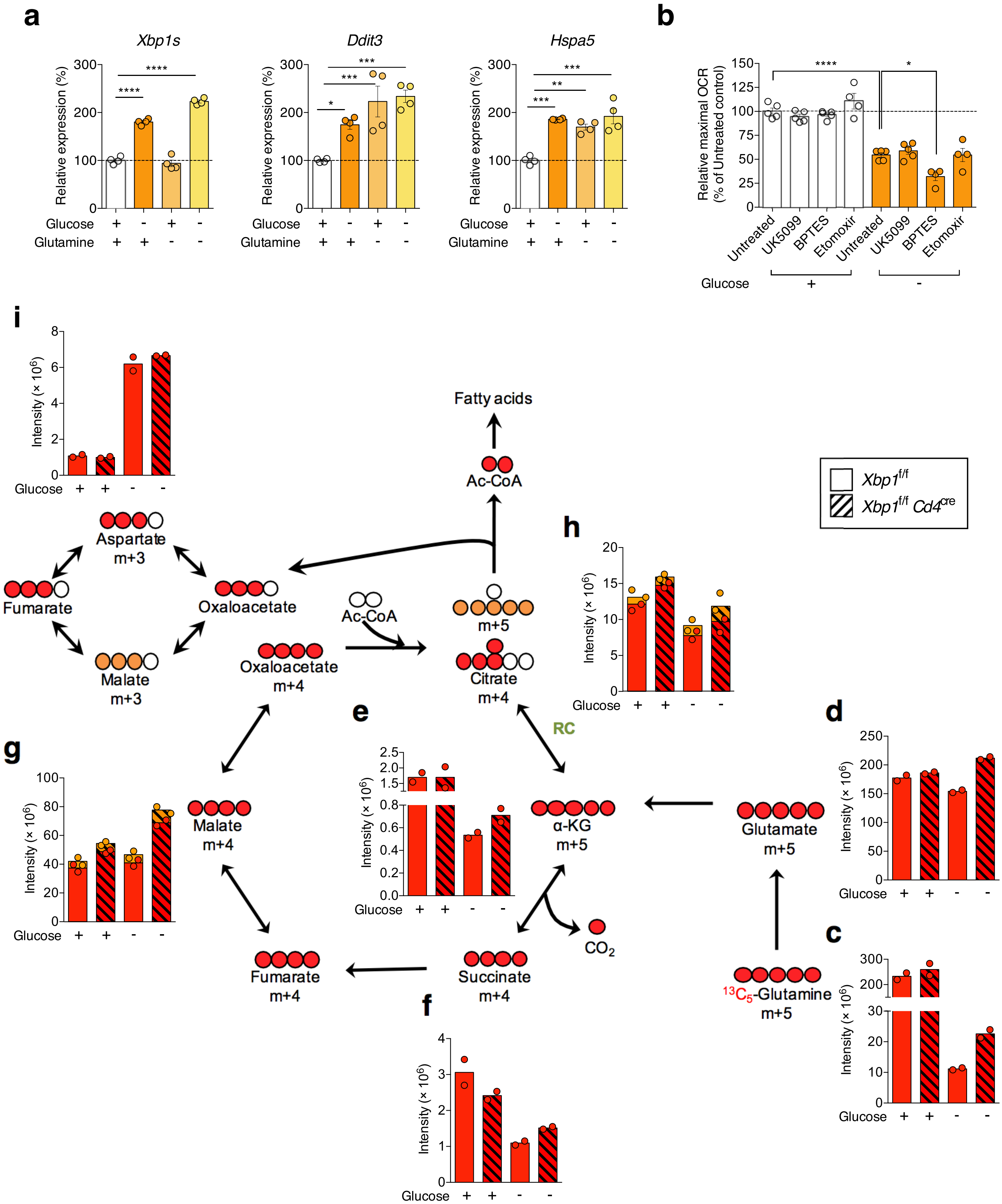
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Reduced glucose uptake and defective *N*-linked protein glycosylation promotes IRE1α–XBP1 activation in human CD4$^+$ T cells exposed to ascites. **a**, Human CD4$^+$ T cells were activated by CD3 and CD28 stimulation for 16 h in the absence or presence of supernatants from ovarian cancer ascites at the indicated concentrations. *XBP1* expression was determined by qPCR (10%, $n = 58$; 50%, $n = 25$; 100%, $n = 30$). Data were normalized to endogenous expression of *ACTB* in each sample. Results are presented as per cent increase in expression compared with untreated controls. **b**, Histograms for FACS-based XBP1s staining in CD4$^+$ T cells treated as indicated. Tm, tunicamycin; 4μ8C, inhibitor of the IRE1α RNase domain. Data were validated by three independent experiments. **c**, **d**, CD4$^+$ T cells were treated with vitamin E (VitE, 50 μM) or vehicle (ethanol, 0.1%) for 1 h and then stimulated with anti-CD3 and anti-CD28 beads for 16 h in the presence of ascites. Intracellular reactive oxygen species staining by DCFDA (**c**) and *XBP1* expression (**d**) ($n = 10$). Data are expressed as per cent response change compared with untreated controls. **e**, Glucose concentration in regular culture medium (cRPMI) and in seven independent samples of ovarian cancer ascites. Each dot represents the mean of two measurements. **f**, *XBP1* expression in the samples described in **a** are displayed, based on the final glucose concentration in the culture after addition of ascites. Three independent ascites samples were used: A10 (triangles), A15 (circles) and A17 (squares) at three different concentrations (10, 50 and 100%). **g**, GLUT1 surface expression on the indicated cell types present in ovarian cancer ascites from six independent patients was determined by GLUT1.RBD staining ($n = 6$). Quantitative analysis (left) and representative histograms (right). **h**, Glucose uptake and XBP1s protein expression in activated CD4$^+$ T cells exposed to three different ascites samples at 10 and 100% for 16 h ($n = 14$). Results are presented as relative to untreated controls. **i**, CD4$^+$ T cells were activated with anti-CD3 and anti-CD28 beads in the presence or absence of ascites, for 16 h. Cells were lysed and the enriched glycoprotein fractions were analysed for *N*-linked glycosylation events by LC–MS/MS. Total ion chromatograms for *N*-glycosylation at amino acid 315 in DPP2 are shown. Numbers in blue indicate abundance of each glycan through quantification of the corresponding peak area. Data are representative of two independent experiments with similar results. **j**, CD4$^+$ T cells were treated with 10 mM GlcNAc for 1 h and stimulated by CD3 and CD28 for 16 h in the presence of 10% ascites. Quantitative analyses for XBP1s protein by FACS (left, $n = 6$) and *XBP1* gene expression by qPCR (right, $n = 15$) are presented as per cent response change compared with untreated controls. **k**, Relative basal and maximal OCR for CD4$^+$ T cells exposed to 10% ascites analysed in Fig. 2g ($n = 16$ total from five independent experiments). Data are expressed as per cent response change compared with untreated controls. **l**, **m**, CD4$^+$ T cells were activated by CD3 and CD28 stimulation for 16 h in the absence or presence of tunicamycin (Tm) at the indicated concentrations ($n = 3$). **l**, *XBP1* expression was determined by qPCR. **m**, OCR profile of Tm-treated CD4$^+$ T cells are shown relative to the untreated control. **n**, **o**, Relative quantification of basal (left) and maximal (right) OCR (**n**), and ECAR measurements (**o**) in all independent samples analysed in Fig. 2i ($n = 9$ total from three independent experiments). **p**, **q**, Relative quantification of basal (left) and maximal (right) OCR (**p**), and ECAR measurements (**q**) for the specimens described in Fig. 2j. ($n = 5$ total from two independent experiments). Data are presented as relative expression compared with matching controls that were not exposed to ascites. Data are shown as mean ± s.e.m. (**a**, **c**, **d**, **f**, **g**, **k**–**m**). *n* values represent biologically independent samples (**a**, **c**–**h**, **j**–**q**). One-way ANOVA with Bonferroni's multiple comparisons test (**a**, **l**); two-tailed Student's *t*-test (**c**, **k**); one-way ANOVA with Tukey's multiple comparisons test (**d**, **g**); two-tailed paired Student's *t*-test (**j**, **n**–**q**); nonparametric Spearman's rank correlation test, Spearman coefficient (r) with P value (two-tailed); 95% confidence interval −0.8811 to −0.1626 (**h**). $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. NS, not significant; MFI, mean fluorescence intensity; gMFI, geometric mean fluorescence intensity.



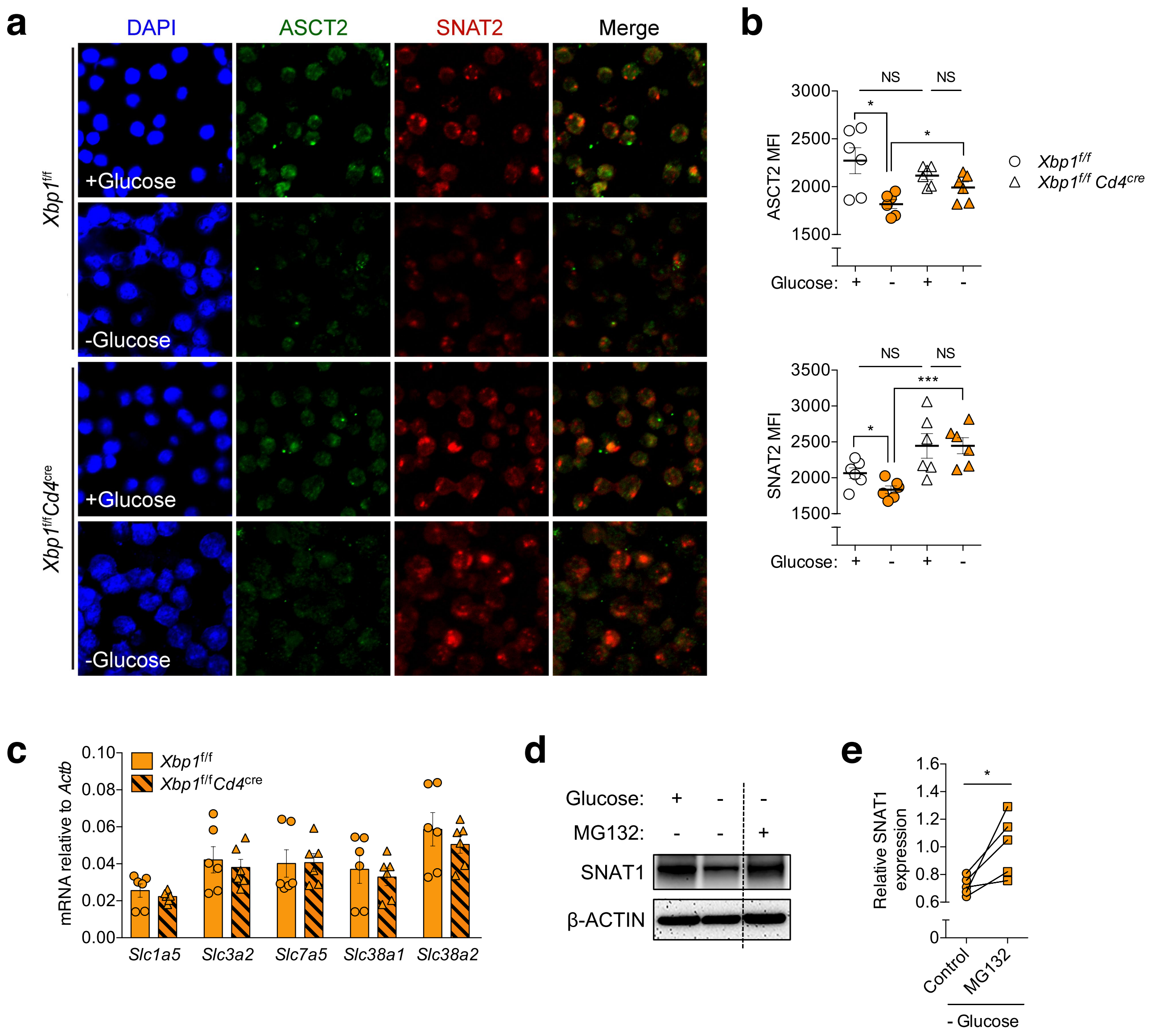
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Characterization of mice devoid of XBP1 in T cells. **a**, Deletion efficiency was analysed by qPCR using a primer set that specifically detects the exon 2 region of *Xbp1*. Data were normalized to endogenous expression of *Actb* and presented as relative expression compared with wild-type littermates ($n=8$). **b**, Absolute cell numbers in the thymus, spleen and lymph nodes. **c**, FACS-based phenotyping of double-negative ($CD4^{-}CD8^{-}$), double-positive ($CD4^{+}CD8^{+}$) or single-positive ($CD4^{+}$ or $CD8^{+}$) thymocytes. **d**–**g**, Frequency of $TCR\beta^{+}$ cells (**d**, **f**) and $CD4^{+}$ or $CD8^{+}$ cells (gated on $TCR\beta^{+}$ cells) (**e**, **g**) in lymph nodes or spleen. **h**, **i**, Expression of CD44 and CD62L on both $CD4^{+}$ (**h**) and $CD8^{+}$ (**i**) $TCR\beta^{+}$ subsets in the spleen. **j**, Frequency of splenic $TCR\beta^{+}CD4^{+}FoxP3^{+}$ T cells. **k**, **l**, Frequency of non-T-cell populations among total live cells in spleen (**k**) and lymph nodes (**l**). **b**–**g**, $n=5$; **h**–**j**, $n=3$; **k**, **l**, *Xbp1$^{f/f}$* ($n=4$), *Xbp1$^{f/f}$Cd4^{cre}* ($n=5$). **m**, Reconstitution efficiency of $CD4^{+}$ and $CD8^{+}$ T cells in bone marrow and spleen from mixed bone marrow chimaeras ($n=3$ per chimaera type). Chimaeras were generated with a mixture of wild-type bone marrow ($CD45.1^{+}$) plus either *Xbp1$^{f/f}$* or *Xbp1$^{f/f}$Vav1^{cre}* bone marrow ($CD45.2^{+}$). **n**, Flow cytometry assessing cell proliferation of $CD4^{+}$ T cells stained with the division-tracking dye (Cell Trace Violet). Cells were left unstimulated or stimulated for 72 and 96 h with plate-bound anti-CD3 (5 μg ml^{-1}) and soluble anti-CD28 (1 μg ml^{-1}). Histograms (left) and proliferation index (right) are shown ($n=4$). **o**, Cell-cycle analysis of $CD4^{+}$ T cells activated for 72 h by staining with propidium iodide. Representative plots from two experiments. **p**, Transmission electron microscopy of in vitro activated wild-type versus XBP1-deficient $CD4^{+}$ T cells. Naive $CD4^{+}$ T cells isolated from three biologically independent mice were activated with plate-bound anti-CD3 and soluble anti-CD28 antibodies for 48 h. White arrowheads indicate the ER; M, mitochondria; magnification 12,000× (left); 50,000× (right). Average mitochondrial area of independent cells was estimated using ImageJ software. *Xbp1$^{f/f}$* ($n=19$); *Xbp1$^{f/f}$Cd4^{cre}* ($n=29$). **q**, Histogram (left) and quantification (right) for mitochondrial staining (Mitotracker) in in vitro activated $CD4^{+}$ T cells ($n=2$ from two independent experiments). **r**, Activated wild-type versus XBP1-deficient $CD4^{+}$ T cells were incubated in glucose-containing, glucose-depleted or 2-deoxyglucose (2-DG, 10 mM)-treated medium for 6 h and PGC1α expression was analysed by immunoblot. β-Actin was used as loading control. Representative plots from two independent experiments. Data are shown as mean ± s.e.m. (**a**–**n**, **p**). *n* values represent biologically independent samples (**a**–**n**, **p**, **q**). Two-tailed Student's *t*-tests (**b**–**l**, **n**, **p**); $^{*}P<0.05$; NS, not significant.

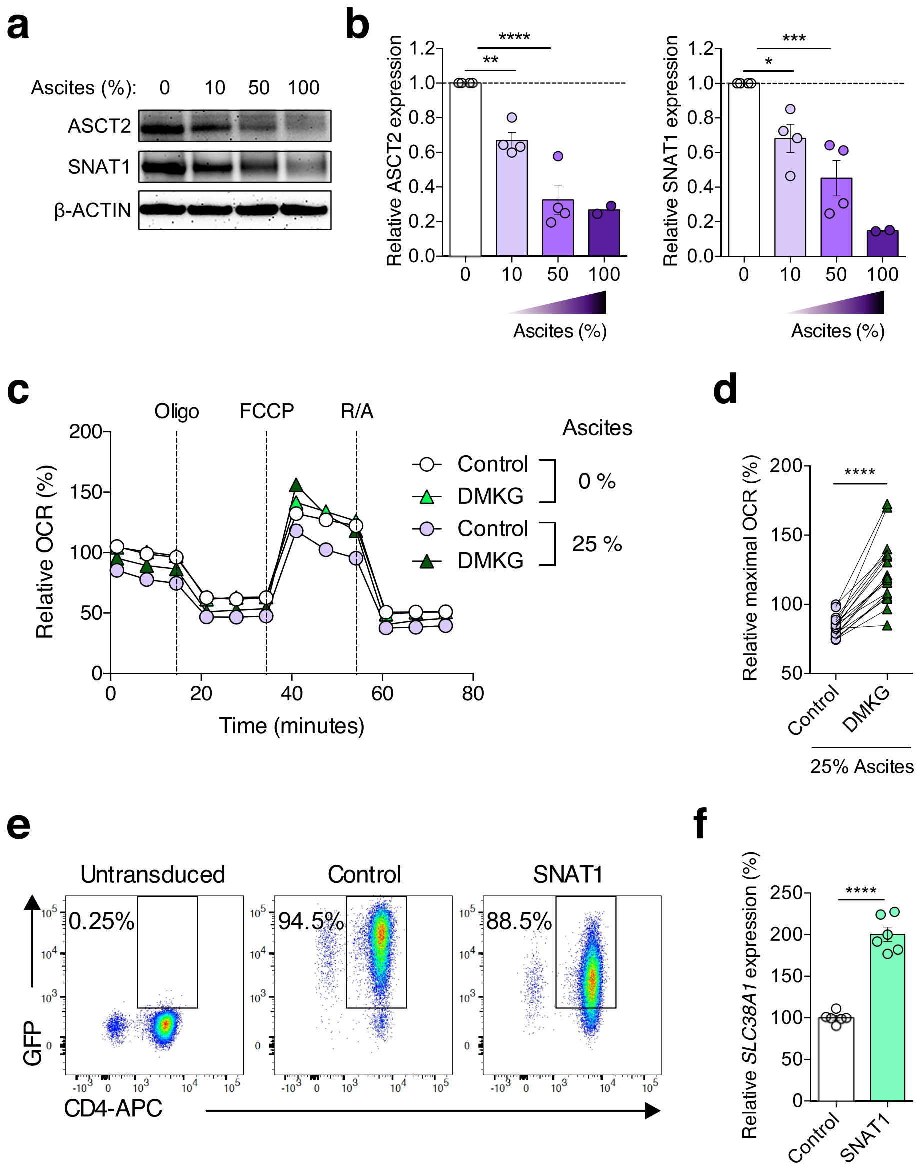


Extended Data Fig. 3 | See next page for caption.

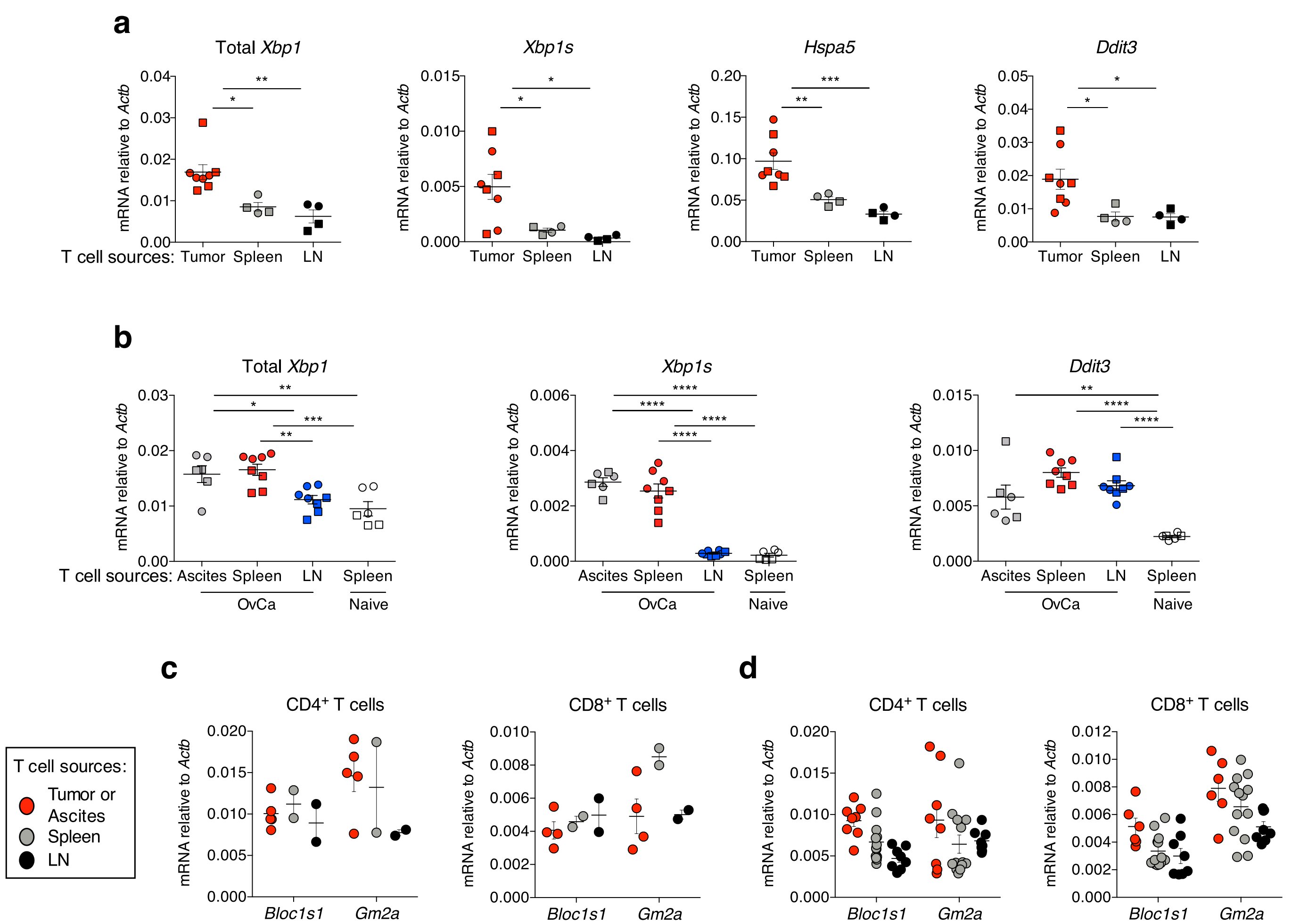
Extended Data Fig. 3 | XBP1 inhibits glutamine influx in response to glucose deprivation. **a**, **b**, Naive splenic CD4$^+$ T cells isolated from wild-type mice were activated by CD3 and CD28 stimulation for 48 h and then incubated for 6 h in the indicated medium. **a**, Expression of gene markers related to ER stress ($n = 4$ from two independent experiments). Data are shown as per cent response change compared with control in the presence of medium containing glucose and glutamine. **b**, Maximal OCR was measured in CD4$^+$ T cells in the presence or absence of glucose, and treated with corresponding medium (untreated, $n = 5$) or inhibitors blocking pyruvate (UK5099, $n = 5$), glutamine (BPTES, $n = 4$) or fatty acid (etomoxir, $n = 4$) oxidation. Data are presented as per cent response change compared with untreated control in the presence of glucose. **c**–**i**, Naive splenic CD4$^+$ T cells isolated from wild-type (solid bars) or XBP1-deficient (hatched bars) mice were activated by CD3 and CD28 stimulation for 48 h, followed by culture in the presence or absence of glucose for 4.5 h, and then pulsed with [U-^{13}C]glutamine for an additional 1.5 h in the same culture condition. Relative abundance of ^{13}C-labelled metabolites and TCA intermediates including glutamine (**c**), glutamate (**d**), α-ketoglutarate (**e**), succinate (**f**), malate (**g**), citrate (**h**) and aspartate (**i**) was determined by LC–MS/MS. Data were normalized to cell number in all cases and are representative of two independent experiments with $n = 2$ biologically distinct samples per group. Data are shown as mean ± s.e.m. n values represent biologically independent samples (**a**, **b**). One-way ANOVA with Bonferroni's multiple comparisons test (**a**); one-way ANOVA with Tukey's multiple comparisons test (**b**); $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.



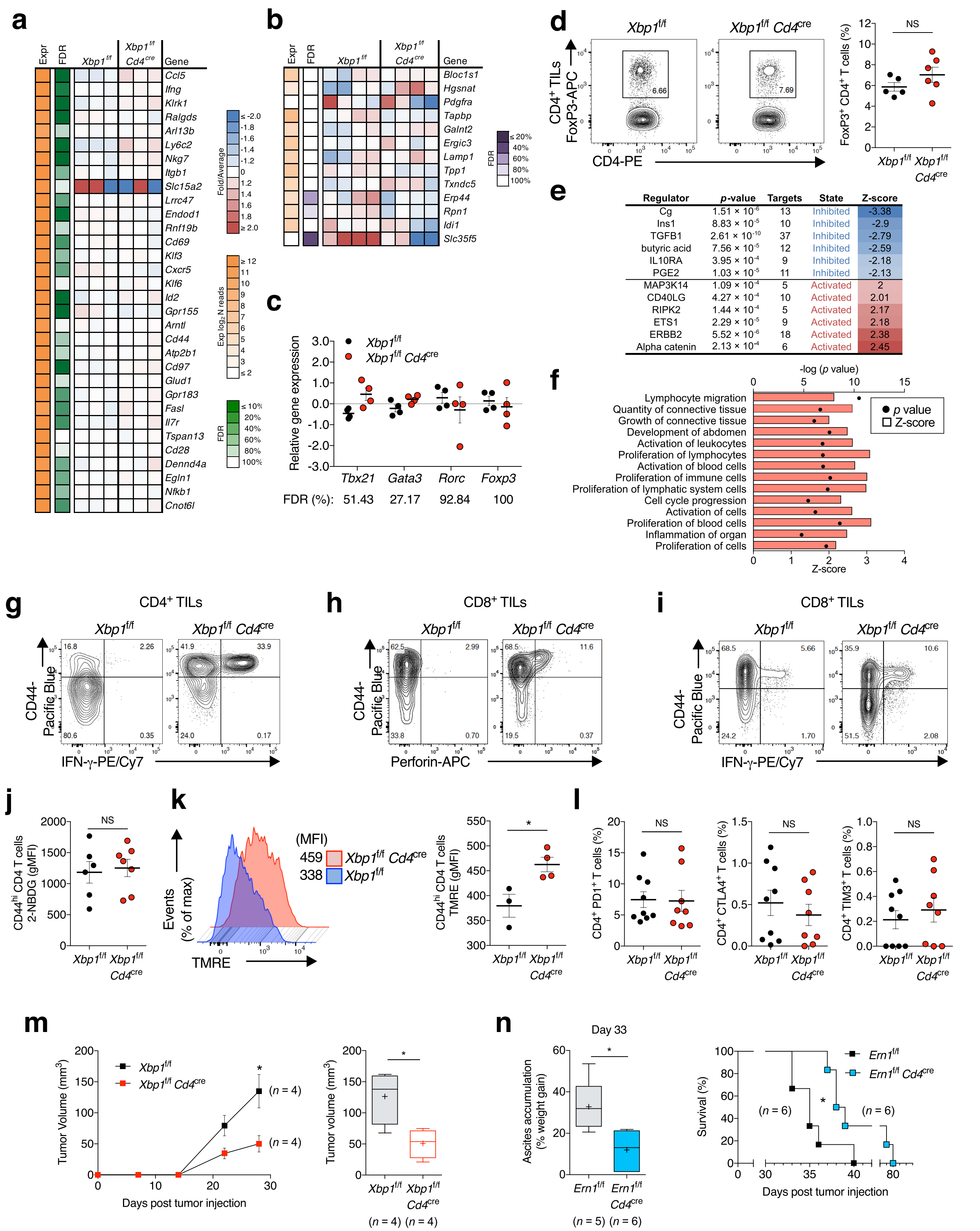
Extended Data Fig. 4 | XBP1 controls the abundance of glutamine transporters in glucose-deprived T cells. **a**, **b**, Pre-activated wild-type or XBP1-deficient CD4$^+$ T cells were incubated in the indicated medium for 6 h and then stained on poly-L-lysine coated discs using antibodies specific for ASCT2 (green) or SNAT2 (red). Nuclei are depicted in blue (DAPI staining). **a**, Representative confocal images of the indicated T cells from three experiments. **b**, The mean fluorescence intensity (MFI) of each glutamine transporter on about 50 individual cells from three independent slides ($n = 150$) was computationally quantified using the ImageJ software by two independent investigators in a blinded manner. Individual dots depict the average MFI of each independent analysis ($n = 6$). **c**, Naive splenic CD4$^+$ T cells isolated from wild-type or XBP1-deficient mice were activated by CD3 and CD28 stimulation for 48 h and then incubated in medium that lacks glucose for 6 h. mRNA expression of genes encoding glutamine transporters was determined by qPCR ($n = 6$ from three experiments). Data were normalized to endogenous expression of *Actb* in each case. **d**, **e**, Pre-activated mouse CD4$^+$ T cells were incubated in the indicated medium for 6 h in the presence or absence of proteasome inhibitor MG132 (10 μM). **d**, Protein levels of the glutamine transporter SNAT1 were determined by immunoblot analysis, in which β-actin was used as loading control. Representative image from five independent experiments. **e**, Densitometric quantification of SNAT1 ($n = 5$). Results are presented as relative expression compared with untreated control T cells incubated in glucose-containing medium. Data are shown as mean ± s.e.m. (**b**, **c**). *n* values represent biologically independent samples (**c**, **e**). Two-tailed Student's *t*-tests (**b**); two-tailed paired Student's *t*-tests (**e**); $^*P < 0.05$, $^{***}P < 0.001$; NS, not significant.



Extended Data Fig. 5 | Restoring glutamine influx enhances mitochondrial function in CD4+ T cells exposed to ascites.
a, **b**, Immunoblot (**a**) and densitometric quantification (**b**) of levels of ASCT2 and SNAT1 protein in human CD4+ T cells exposed to ovarian cancer ascites at the indicated concentrations for 16 h. β-Actin was used as loading control. Data are shown as the relative expression compared with untreated (0%) controls. $n = 4$ for 10% ascites; $n = 4$ for 50% ascites; $n = 2$ for 100% ascites. Data were generated from two independent experiments. **c**, **d**, Human CD4+ T cells were activated by CD3 and CD28 stimulation for 16 h in the absence or presence of 25% supernatants of ovarian cancer ascites, and DMKG (5 mM) was added to the cell culture during the last 4 h of incubation. OCR profile (**c**) and quantification of maximal OCR (**d**). Data are presented as relative expression compared with untreated controls incubated in the absence of ascites ($n = 17$ total from two independent experiments). **e**, **f**, Human CD4+ T cells activated by CD3 and CD28 stimulation and IL-2 (50 U ml^{-1}) for 36 h were transduced with GFP-expressing retroviruses containing no insert (control) or the gene that encodes human SNAT1. GFP+ cells were sorted 3 days after transduction and expanded for an additional 48 h in the presence of CD3 and CD28 stimulation, and IL-2 (50 U ml^{-1}). After 20 h of resting, cells were restimulated with CD3 and CD28 antibodies in the absence or presence of supernatants of ovarian cancer ascites for 16 h. **e**, Sorting plots showing GFP expression by CD4+ T cells that were left untreated or transduced with either control or SNAT1-expressing viruses. Representative plots from two experiments. **f**, Relative *SLC38A1* expression levels in sorted cells after transduction ($n = 6$ total from two independent experiments). n values represent biologically independent samples (**b**, **d**, **f**). Data are shown as mean ± s.e.m. (**b**, **c**, **f**). One-way ANOVA with Bonferroni's multiple comparisons test (**b**); two-tailed paired Student's t-test (**d**); two-tailed Student's t-test (**f**). $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$.

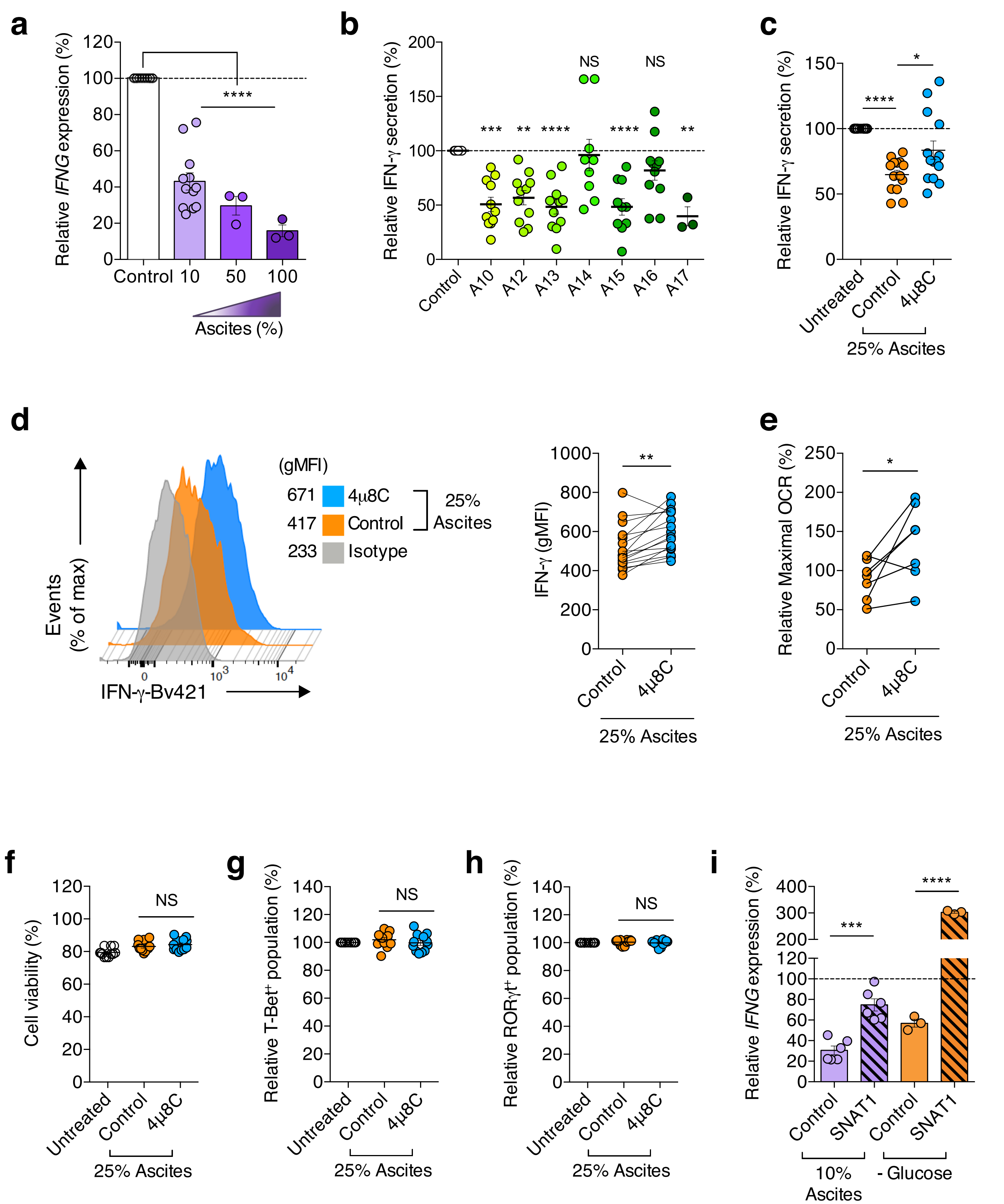


Extended Data Fig. 6 | IRE1α–XBP1 activation and ER stress responses in ovarian cancer-associated T cells isolated from mouse models of disease. a, **b**, Expression of marker genes (*Xbp1*, *Xbp1s*, *Hspa5* and *Ddit3*) for ER stress was determined by qPCR. Circles, CD4$^+$ T cells; squares, CD8$^+$ T cells. **a**, CD45$^+$TCRβ$^+$CD4$^+$ and CD45$^+$TCRβ$^+$CD8$^+$ cells were sorted from tumours ($n = 8$), spleens ($n = 4$) and lymph nodes ($n = 4$) of mice bearing advanced ovarian tumours driven by p53 and KRAS. **b**, CD45$^+$TCRβ$^+$CD4$^+$ and CD45$^+$TCRβ$^+$CD8 cells were isolated from malignant ascites ($n = 6$), spleens ($n = 8$) and lymph nodes ($n = 8$) of mice bearing aggressive ID8-*Defb29*/*Vegfa* ovarian cancer (OvCa), and from spleens ($n = 6$) of naive mice as control. **c**, **d**, Expression of target genes (*Bloc1s1* and *Gm2a*) of canonical regulated IRE1α-dependent decay (RIDD)—in CD4$^+$ and CD8$^+$ T cells isolated from different tissues of mice that bore ovarian cancer—was analysed by qPCR. **c**, T cells were sorted from tumours (CD4$^+$ T cells, $n = 5$; CD8$^+$ T cells, $n = 4$), spleens ($n = 2$) and lymph nodes ($n = 2$) of mice bearing advanced ovarian tumours driven by p53 and KRAS. **d**, T cells were isolated from malignant ascites (CD4$^+$ T cells, $n = 8$; CD8$^+$ T cells, $n = 6$), spleens ($n = 13$) and lymph nodes ($n = 8$) of mice that bore aggressive ID8-*Defb29*/*Vegfa* ovarian cancer. Data were normalized to *Actb*. Data are shown as mean ± s.e.m. (**a**–**d**). *n* values represent biologically independent samples (**a**–**d**). One-way ANOVA with Tukey's multiple comparisons test (**a**, **b**); $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.



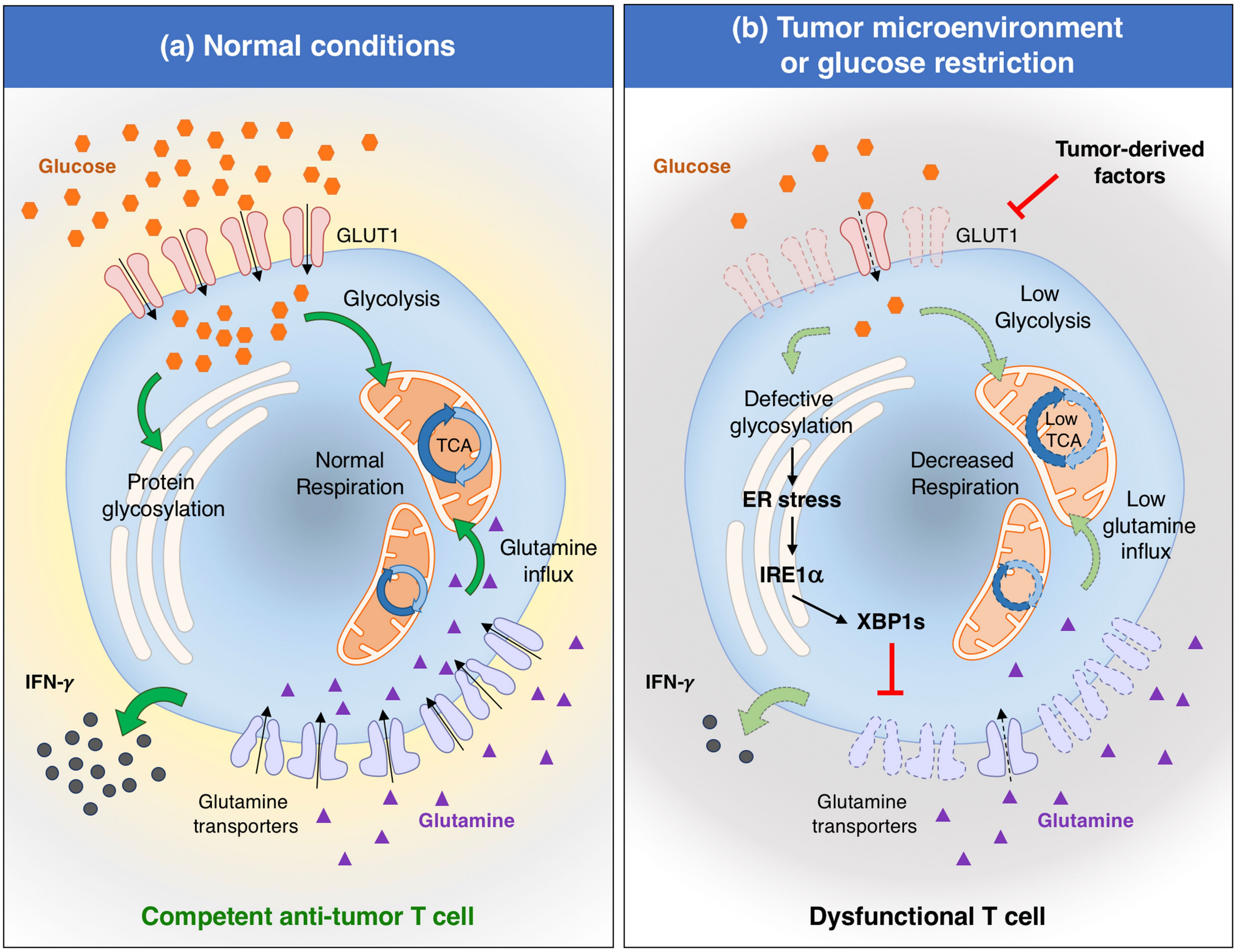
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | IRE1α–XBP1 signalling alters function of T cells associated with ovarian cancer, and promotes malignant progression. **a**, Transcriptional profiling of splenic CD44^{high}CD62L^{low}CD4^{+} T cells sorted from naive wild-type versus XBP1-deficient mice. Expression of the differentially regulated genes identified in Fig. 4a is shown ($n = 3$ per group). **b**, **c**, **e**, **f**, Analysis of wild-type versus XBP1-deficient CD4^{+} T cells isolated from mice that bore metastatic ovarian cancer ($n = 4$ per group). **b**, Expression of previously reported target genes of regulated IRE1α-dependent mRNA decay (RIDD). **c**, Relative gene expression of master transcription factors controlling helper T cell differentiation. **d**, Intracellular staining for FoxP3 (left) and proportion of FoxP3^{+}CD4^{+} T cells from wild-type ($n = 5$) and XBP1-deficient ($n = 6$) mice that bore metastatic ovarian cancer for 21 days. **e**, Predicted upstream regulators associated with the transcriptional changes observed. **f**, Enriched cellular functions in XBP1-deficient CD4^{+} T cells at tumour sites. *Z*-scores greater than 2 indicate functions predicted to be significantly increased in XBP1-deficient CD4^{+} T cells. **g**, Intracellular staining for IFNγ in CD45^{+}CD3^{+}CD4^{+} T cells from wild-type and conditional XBP1-deficient mice that bore metastatic ovarian cancer for 29 days (late stage). Representative plots from three independent experiments. **h**, Intracellular staining for perforin in CD45^{+}CD3^{+}CD8^{+} T cells from wild-type and conditional XBP1-deficient mice that bore metastatic ovarian cancer for 23 days. **i**, Intracellular staining for IFNγ in CD45^{+}CD3^{+}CD8^{+} T cells from wild-type and conditional XBP1-deficient mice that bore metastatic ovarian cancer for 29 days (late stage). Representative plots from two independent experiments. **j**, In vivo glucose uptake by CD44^{high}CD4^{+} tumour-infiltrating lymphocytes in *Xbp1$^{f/f}$* ($n = 6$) or *Xbp1$^{f/f}$Cd4^{cre}* ($n = 7$) female mice that bore metastatic ovarian cancer. **k**, Representative TMRE staining for CD45^{+}TCRβ^{+}CD44^{+}CD4^{+} T cells associated with ovarian cancer, from tumour-bearing *Xbp1$^{f/f}$* ($n = 3$) or *Xbp1$^{f/f}$Cd4^{cre}* ($n = 4$) mice. **l**, Peritoneal wash samples were collected from mice at 24 days after tumour challenge and the proportion of CD4^{+} T cells associated with ovarian cancer that expressed PD-1, CTLA4 and TIM3 in wild-type ($n = 9$) or XBP1-deficient ($n = 8$) mice was determined. **m**, Female mice ($n = 4$ per group) were implanted with ID8-*Defb29/Vegfa* ovarian cancer cells in the flank, and tumour growth was monitored over time (left). Tumours were resected at day 34 and final size was confirmed ex vivo (right). **n**, Ascites accumulation (left) in *Ern1$^{f/f}$* ($n = 5$) or *Ern1$^{f/f}$Cd4^{cre}* ($n = 6$) mice that bore ID8-*Defb29/Vegfa* ovarian cancer and overall survival (right, $n = 6$ per group). *Ern1* is the gene that encodes IRE1α. Data are shown as mean ± s.e.m. (**c**, **d**, **j**–**l**). Boxes represent median ± interquartile range and whiskers indicate minimum and maximum (**m**, **n**). *n* values represent biologically independent mice (**a**–**d**, **j**–**n**). Two-tailed Student's *t*-tests (**d**, **j**–**l**, **m**, **n**); log-rank test for survival (**n**). *$P < 0.05$, NS, not significant; gMFI, geometric mean fluorescence intensity.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | IRE1α–XBP1 regulates IFNγ production in human CD4$^+$ T cells exposed to ascites. **a**, *IFNG* expression in CD4$^+$ T cells receiving CD3 and CD28 activation for 16 h under increasing concentrations of supernatants of ovarian cancer ascites; 10% ($n = 12$); 50% ($n = 3$); 100% ($n = 3$). **b**, CD4$^+$ T cells were activated for 12 h, incubated for additional 36 h in the presence of 25% ascites, and IFNγ in culture supernatants was determined by ELISA. Data were normalized to final viable cell counts in each case. $n = 11$ independent responder CD4$^+$ T cells in all cases with the exception of A14 ($n = 9$), A15 ($n = 10$) and A17 ($n = 3$). **c**–**h**, Pre-activated CD4$^+$ T cells were treated with 4μ8C for 2 h, and 25% ascites was subsequently added to the culture for additional 12 h. **c**, IFNγ in culture supernatants was quantified by ELISA ($n = 14$). Data are presented as relative expression compared with matching controls that were not exposed to ascites. **d**, FACS histogram (left) and quantitative analysis (right) for intracellular IFNγ in CD4$^+$ T cells ($n = 16$). **e**, Maximal OCR presented as per cent response change compared with untreated controls ($n = 7$). **f**, The frequency of viable cells among total cells was determined by staining with dead cell exclusion dye ($n = 12$). The frequency of T-bet$^+$ (**g**) and RORγt$^+$ (**h**) populations among CD4$^+$ T cells was determined by intracellular staining and presented as relative expression compared with controls that were not treated with ascites ($n = 12$). **i**, *IFNG* expression by SNAT1-overexpressing human CD4$^+$ T cells exposed to 10% ovarian cancer ascites ($n = 6$ from two independent experiments) or incubated in glucose-free medium ($n = 3$ from two independent experiments). Data were normalized to endogenous expression of *GAPDH* in each sample. Data are presented as relative expression compared with control-virus-transduced T cells that were not exposed to ascites or glucose-deprived. Data are shown as mean ± s.e.m. (**a**–**c**, **f**–**i**). n values represent biologically independent samples (**a**–**i**). One-way ANOVA with Bonferroni's multiple comparisons test (**a**, **b**); one-way ANOVA with Tukey's multiple comparisons test (**c**, **f**–**h**); two-tailed paired Student's t-tests (**d**, **e**); two-tailed Student's t-tests (**i**); $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$; gMFI, geometric mean fluorescence intensity.



Extended Data Fig. 9 | Proposed model. a, Under normal (glucose-rich) conditions, T cells can effectively glycosylate proteins in the ER while fuelling mitochondrial respiration through glycolysis. These processes endow T cells with competent effector function and anti-cancer capacity. **b**, In the tumour microenvironment, glucose availability could be limited and T cells may also express low levels of glucose transporters such as GLUT1. Glucose restriction not only dampens glycolysis, but also impairs optimal *N*-linked protein glycosylation in T cells, leading to ER stress and IRE1α–XBP1 activation. XBP1 controls the abundance of glutamine transporters in T cells that are experiencing ER stress, and consequently limits the influx of glutamine necessary to sustain mitochondrial respiration under glucose deprivation. Therefore, T cells become dysfunctional and incapable of controlling malignant progression. Disabling IRE1α–XBP1 signalling may be useful to enhance T cell mitochondrial function and anti-cancer capacity in a harsh tumour microenvironment.

Extended Data Table 1 | Defective *N*-linked protein glycosylation in $CD4^+$ T cells exposed to ascites

Protein name	N-linked Glycosylation Site	N-linked Glycosylation Glycan	Number of glycosylation events: Control	Number of glycosylation events: Ascites	% decrease in glycosylation
CD48	207	HexNAc(2)Hex(5)	1	1	25
		HexNAc(2)Hex(6)	2	1	
		HexNAc(3)Hex(6)NeuAc(1)	1	1	
DPP2 (Dipeptidyl peptidase 2)	315	HexNAc(2)Hex(2)	1	0	18.2
		HexNAc(2)Hex(3)	2	2 (*)	
		HexNAc(2)Hex(4)	1	2	
		HexNAc(2)Hex(5)	1	2 (*)	
		HexNAc(2)Hex(6)	1	0	
	428	HexNAc(2)Hex(5)	3	1	
		HexNAc(2)Hex(6)	2	2 (*)	
HEXB (β-hexosaminidase B)	84	HexNAc(2)Hex(5)	2	0	75
		HexNAc(2)Hex(6)	2	1	
MPO (Myeloperoxidase)	323	HexNAc(2)Hex(3)	4	0	79.5
		HexNAc(2)Hex(4)	5	0	
		HexNAc(2)Hex(5)	2	0	
		HexNAc(2)Hex(6)	4	0	
		HexNAc(2)Hex(7)	2	0	
		HexNAc(2)Hex(2)Fuc(1)	3	0	
		HexNAc(4)Hex(3)NeuGc(1)	1	0	
	355	HexNAc(2)Hex(4)	2	0	
		HexNAc(2)Hex(5)	4	3 (*)	
		HexNAc(2)Hex(6)	7	3	
		HexNAc(2)Hex(7)	3	2	
	391	HexNAc(2)Hex(3)	1	0	
		HexNAc(2)Hex(4)	3	0	
		HexNAc(2)Hex(5)	2	0	
		HexNAc(2)Hex(6)	1	1	

$CD4^+$ T cells were activated with anti-CD3 and anti-CD28 beads in the presence or absence of ascites for 16 h. Cells were lysed and the enriched glycoprotein fractions were analysed for *N*-linked glycosylation events by LC–MS/MS. Representative glycoproteins recovered from both control and ascites-exposed $CD4^+$ T cells. The table shows each site for *N*-linked glycosylation on the identified protein, the glycan type on that site and the number of glycosylation events identified in each glycoform. The per cent decrease in glycosylation upon ascites exposure was calculated using the following equation: per cent decrease in glycosylation = 100 – (number of total glycosylation events$_{ascites}$ divided by the number of total glycosylation events$_{control}$) × 100.
*Altered event.

https://doi.org/10.1038/s41586-018-0596-y

A cell identity switch allows residual BCC to survive Hedgehog pathway inhibition

Brian Biehs[1], Gerrit J. P. Dijkgraaf[1], Robert Piskol[2], Bruno Alicke[3], Soufiane Boumahdi[1], Franklin Peale[4], Stephen E. Gould[3] & Frederic J. de Sauvage[1]*

Despite the efficacy of Hedgehog pathway inhibitors in the treatment of basal cell carcinoma (BCC)[1], residual disease persists in some patients and may contribute to relapse when treatment is discontinued[2]. Here, to study the effect of the Smoothened inhibitor vismodegib on tumour clearance, we have used a *Ptch1–Trp53* mouse model of BCC[3] and found that mice treated with vismodegib harbour quiescent residual tumours that regrow upon cessation of treatment. Profiling experiments revealed that residual BCCs initiate a transcriptional program that closely resembles that of stem cells of the interfollicular epidermis and isthmus, whereas untreated BCCs are more similar to the hair follicle bulge. This cell identity switch was enabled by a mostly permissive chromatin state accompanied by rapid Wnt pathway activation and reprogramming of super enhancers to drive activation of key transcription factors involved in cellular identity. Accordingly, treatment of BCC with both vismodegib and a Wnt pathway inhibitor reduced the residual tumour burden and enhanced differentiation. Our study identifies a resistance mechanism in which tumour cells evade treatment by adopting an alternative identity that does not rely on the original oncogenic driver for survival.

Basal cell carcinoma is the most frequently occurring human cancer and results from aberrant activation of the Hedgehog (Hh) pathway[4]. Erivedge (vismodegib), a potent inhibitor of Hh signalling that acts at the level of Smoothened (SMO)[5], has been approved for the treatment of locally advanced and metastatic BCC[1,6,7], and is also effective for operable BCC[2]. However, the low rate of histological clearance observed raised concerns over whether residual BCCs can regrow once treatment is discontinued.

To determine how residual BCCs survive vismodegib treatment, we used a mouse model of BCC driven by inactivation of *Ptch1* and *Trp53* in skin basal cells[3]. Histological examination of skin from 8-week-old *K14*Cre^{ER};*Ptch1*$^{fl/fl}$;*Trp53*$^{fl/fl}$ mice revealed superficial BCC, similar to the human condition (Fig. 1a, b). These tumours expressed high levels of the Hh target gene *Gli1* and the BCC marker SOX9[8] (Fig. 1f, g). Ki67 staining revealed that roughly 50% of the tumour nests contained actively proliferating cells (Fig. 1h).

To test the effect of Hh pathway inhibition on established BCC, we treated 8 week-old mice with vismodegib for 28 days. Skin from treated mice frequently contained small residual tumours similar to the finger-like basaloid projections observed in patients[2] (Fig. 1c, d). The absence of *Gli1* expression in tumour nests indicated that vismodegib effectively blocked Hh signalling (Fig. 1i). Residual BCCs still expressed SOX9 (Fig. 1j) and stopped proliferating completely within 1 day of treatment initiation (Fig. 1k). Apoptosis was absent in untreated disease, but was observed in a small number of tumour cells early during treatment (Extended Data Fig. 1a–d). Residual tumours did not differentiate into normal epidermis (Extended Data Fig. 2) and remained present even after 56 days of drug treatment (Fig. 1e). Thus, residual BCCs in mice are quiescent and do not require active Hh signalling for their survival.

We next investigated whether residual lesions that survived drug treatment could reinitiate growth. Skin sections from BCC mice collected 0, 3, 6, or 12 days after the last dose revealed that superficial BCC returned quickly (Fig. 2a), with *Gli1*$^{+}$ lesions resembling either residual tumours or newly formed tumour buds (Fig. 2b–d). Elevated *Gli1* expression persisted as mice remained off drug for longer periods of time (Extended Data Fig. 1e, f). To distinguish growth of residual disease from de novo tumour formation due to inherent leakiness[9] of the *K14*Cre^{ER}, we labelled tumours with BrdU before treatment to mark residual disease (Fig. 2e, Extended Data Fig. 1g, h). Six days after cessation of vismodegib, labelled tumour nests stained positive for Ki67 and were diluting the incorporated BrdU, confirming that quiescent residual BCCs reinitiated tumour growth when drug treatment was discontinued (Fig. 2f, g).

To uncover the mechanisms of BCC persistence, we performed RNA sequencing (RNA-seq) and compared the transcriptional profiles of untreated Ki67^{+} and Ki67^{-} BCCs to those of residual tumours. Independent of proliferation status (Extended Data Fig. 3c), vismodegib affected the expression of known Hh targets (Fig. 3a) and a large number of other genes relative to untreated tumours (Extended Data Fig. 3a, b). Comparison with published skin compartment signatures (Extended Data Table 1) revealed that genes enriched in hair follicle bulge stem cells were downregulated, whereas genes enriched in basal cells from the interfollicular epidermis (IFE) or isthmus (ISTH) showed upregulation after treatment (Fig. 3b, Extended Data Fig. 3d–f). There was no significant overlap with hair germ, dermal papillae or outer root sheath (ORS) signatures (Fig. 3b, Extended Data Table 1). To confirm the shift from a hair follicle bulge to a mixed IFE/ISTH identity, we performed in situ hybridization (ISH) using key markers of these skin compartments. The transcription factor *Lhx2*, which has a key role in hair follicle bulge stem cells[10] (Extended Data Fig. 3g), was abundant in untreated BCCs, but strongly depleted in residual disease (Fig. 3c, f, Extended Data Fig. 3l). On the other hand, IL-33 and *Defb6*, which are expressed in the normal IFE and ISTH, respectively (Extended Data Fig. 3h, i), were absent in established BCCs (Fig. 3d, Extended Data Fig. 3j), but induced in residual disease (Fig. 3g, Extended Data Fig. 3k), as was the ISTH marker MTS24[11] (Extended Data Fig. 3l). Similar results for LHX2 and IL-33 were obtained in a lineage-tracing experiment with a Cre-inducible TdTomato reporter (Extended Data Fig. 4a–d), confirming that the cell identity shift occurs in residual tumour cells. Notably, the absence of IL-33 staining in untreated tumours and gradual appearance of this nuclear factor[12] in residual disease (Extended Data Fig. 5) suggests that vismodegib induces a cell identity switch rather than selecting for pre-existing ISTH/IFE-like cells. The shift in cell identity was reversible upon cessation of drug treatment, as the proportion of BrdU^{+} residual tumour cells expressing LHX2 increased progressively in our BrdU labelling experiment (Extended Data Fig. 4e–i). Furthermore, GATA6 and KLF5 are normally restricted in skin to the ISTH and IFE, respectively[13,14]. Both transcription factors were undetectable in untreated BCCs, but were

[1]Department of Molecular Oncology, Genentech, San Francisco, CA, USA. [2]Department of Bioinformatics and Computational Biology, Genentech, San Francisco, CA, USA. [3]Department of Translational Oncology, Genentech, San Francisco, CA, USA. [4]Department of Research Pathology, Genentech, San Francisco, CA, USA. *e-mail: desauvage.fred@gene.com

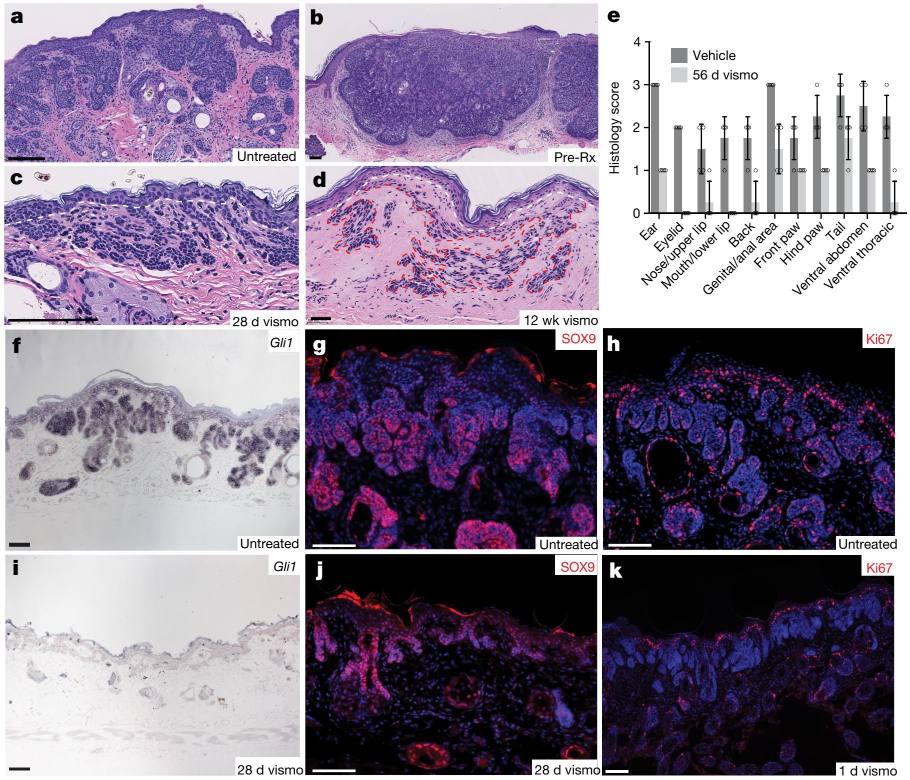


Fig. 1 | BCCs persist in the absence of Hh signalling. a–d, Representative images of haematoxylin and eosin (H&E)-stained skin sections showing response of mouse ($n = 6$ per group) and human BCC ($n = 24$, cohort 1 of ref. [2]) to vismodegib (vismo) treatment. **a**, Skin section from an 8-week-old BCC mouse. **b**, Skin biopsy from a patient with BCC (pre-Rx, pre-treatment). **c**, Residual disease in a BCC mouse. White dashed line denotes boundary between epidermis and dermis. **d**, Residual disease (outlined in red) in the patient shown in **b**. **e**, Histology scores (mean ± s.d.; circles show individual scores) of skin samples from BCC mice treated with vehicle (black) or vismodegib (grey; $n = 4$ per group). **f–k**, Representative images of *Gli1*, SOX9 and Ki67 expression in mouse BCC and residual disease ($n = 4$ per condition). **f**, The Hh pathway is active in mouse BCC. **g**, Mouse BCCs stain positive for SOX9. **h**, Half of the tumours stain positive for Ki67. **i**, The Hh pathway is blocked in residual BCCs during vismodegib treatment. **j**, Residual disease identified by Sox9. **k**, BCCs stop proliferating within 1 day of vismodegib treatment. All experiments were replicated at least twice. Scale bars: **b**, **d**, 90 μm; other panels, 100 μm. DAPI nuclear stain is in blue (**g**, **h**, **j**, **k**); n represents the number of either mice or patients.

co-expressed in residual tumour cells (Fig. 3e, h), suggesting that an induced state of 'lineage infidelity'[14] may contribute to the development of residual disease.

To better define the mechanisms that drive this identity switch, we performed assay for transposase-accessible chromatin with high-throughput sequencing (ATAC–seq) coupled with RNA-seq on tumour cells isolated by fluorescence-activated cell sorting (FACS) before and after either 6 or 14 days of treatment (Extended Data Fig. 6a–f, i). Vismodegib had a relatively small effect on chromatin accessibility, as the majority of the open chromatin regions were shared between treatment groups (Extended Data Fig. 6g). Assessment of the regulatory potential[15] exerted by differentially accessible chromatin on skin compartment signatures did not show enrichment for either opening or closing regions at these loci (Fig. 3i). Instead, these genes were more

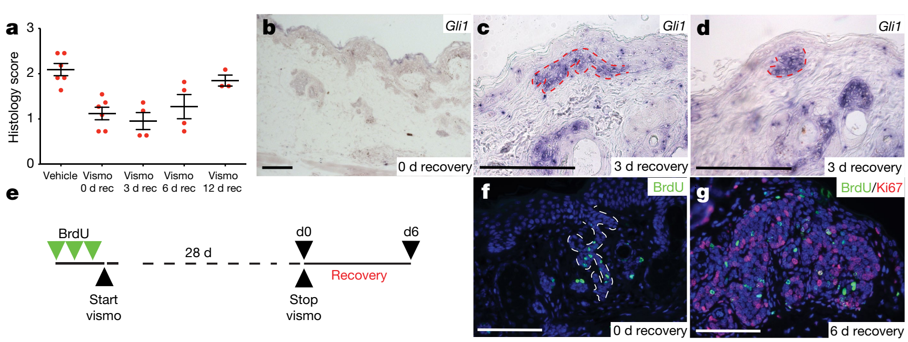


Fig. 2 | Residual BCCs resume growth when vismodegib is discontinued. a, Histology scores (mean ± s.d.) from BCC mice treated with vehicle or vismodegib for 28 days determined after the indicated number of days of recovery (rec) ($n \geq 3$ per group). **b–d**, Representative images of *Gli1* ISH on skin from BCC mice shown in **a**. **b**, Hh pathway is blocked in residual BCC immediately after treatment. **c**, **d**, A residual BCC (**c**, outlined) and a newly formed tumour bud (**d**, outlined), both with an active Hh pathway 3 days after treatment. **e–g**, Lineage tracing of residual BCCs and their subsequent growth after treatment. Representative images of residual BCCs stained for BrdU (green), Ki67 (red) and DAPI nuclear stain (blue) are shown ($n = 4$ per time point). **e**, BrdU labelling strategy of residual BCCs. d0, day 0 of recovery. **f**, $BrdU^+$ residual tumour nests (outlined) are quiescent immediately after treatment. **g**, Residual BCCs have resumed growth 6 days after treatment. All experiments were replicated at least twice. Scale bars are 100 μm; n represents the number of mice.

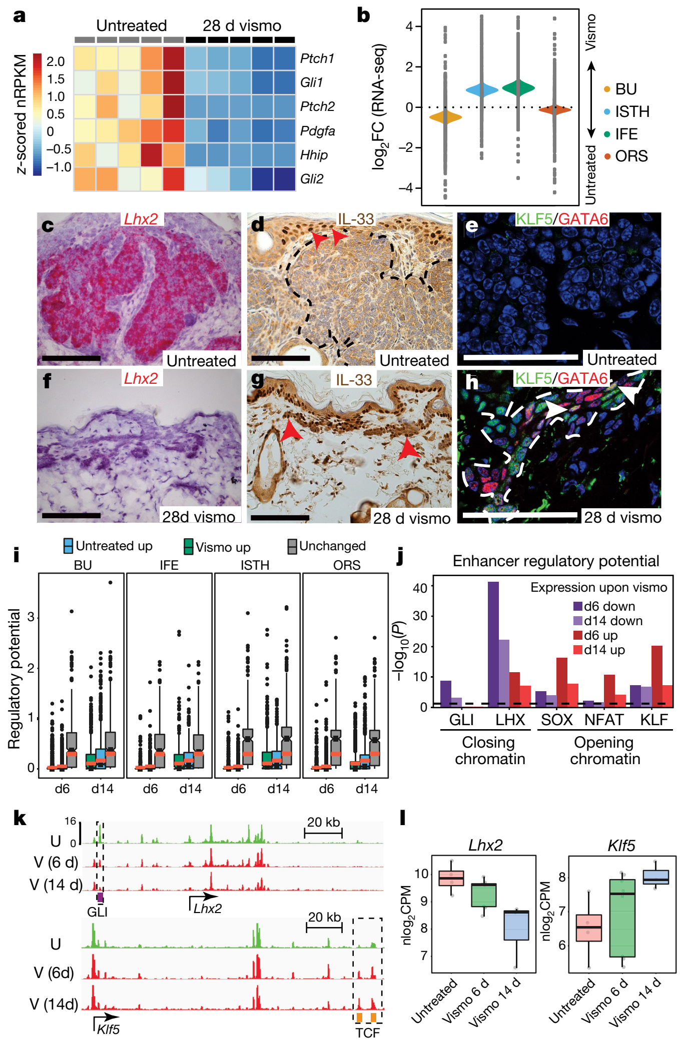


Fig. 3 | Residual BCCs adopt an ISTH/IFE fate upon vismodegib treatment. **a**, Heat-map of Hh target gene expression in untreated and residual BCCs ($n = 5$ per group; data are represented as mean-centred normalized reads per kilobase of transcript per million mapped reads (nRPKM)). **b**, Quantitative set analysis for gene expression (QuSAGE) of indicated signatures in residual versus Ki67^{+} untreated BCC ($n = 5$ per group). Coloured violins depict differential expression of entire gene set; grey dots represent the $\log_2$[fold change] in expression of individual genes. BU, hair follicle bulge. **c**, **d**, **f**, **g**, Representative images of *Lhx2* (hair follicle bulge) and *Il33* (IFE) expression in untreated and residual BCCs ($n = 4$ per condition). **c**, Untreated BCC with high levels of *Lhx2* mRNA. **d**, Untreated BCC (outlined) lacking IL-33 staining. Red arrowheads indicate nuclear localization of IL-33 in the epidermis. **f**, Residual BCCs downregulate *Lhx2*. **g**, Residual BCCs express IL-33 (red arrowheads). **e**, **h**, Representative images of KLF5 (IFE) and GATA6 (ISTH) antibody staining in untreated and residual BCC ($n = 4$ per condition). Untreated BCCs (**e**) lack expression of these lineage-specific transcription factors, whereas residual tumour cells (**h**, outlined) co-express KLF5 (green) and GATA6 (red). **i**, Regulatory potential exerted by chromatin regions (opening, green; closing, blue; unchanged, grey) on indicated gene sets ($n = 3$ per group). Box plots show median, two hinges (25th and 75th percentile), two whiskers (1.5 × inter-quartile range (IQR)), and all outlying points individually. Orange bars indicate the regulatory potential when using random genes. **j**, Graph depicting the association between differentially accessible enhancers with indicated transcription factor binding motifs and differentially expressed genes. A one-tailed Kolmogorov–Smirnov test was used to determine significance ($n = 3$ per group). **k**, Chromatin traces showing mean ATAC peaks at the *Lhx2* and *Klf5* loci. U, untreated; V, vismodegib-treated. **l**, Expression of *Lhx2* and *Klf5* in FACS-sorted BCC cells ($n = 3$ per group; data represented as mean ± s.d. $\log_2$ normalized read counts per million (CPM)). All experiments were replicated at least twice. Scale bars, 100 μm; *n* represents the number of mice.

closely associated with constitutively open chromatin, indicating that in BCC, lineage-specific programs harbour a globally open chromatin structure that is poised for activation, similar to the small intestine[16].

We next assessed whether chromatin accessibility correlated with general changes in gene expression. Closing regions were associated with vismodegib-downregulated genes, whereas opening chromatin was associated with upregulated genes (Extended Data Fig. 6j, k). Closing regions were enriched for GLI, the forkhead box (FOX) family, and LHX binding motifs (Extended Data Fig. 7a), consistent with inhibition of the Hh pathway and the BCC transcriptome resembling the hair follicle bulge compartment. Regions that opened during vismodegib treatment were enriched for binding motifs corresponding to SOX, NFAT, KLF and GATA transcription factors (Extended Data Fig. 7b), consistent with treated BCCs shifting towards an IFE and ISTH identity. Binding and expression target analysis (BETA) confirmed that peaks containing GLI, LHX2, SOX and KLF motifs were strongly associated with changes in gene expression (Fig. 3j, Extended Data Fig. 7c). Notably, TCF binding motifs were enriched in opening chromatin at the 6-day time point (Extended Data Fig. 7b), including the loci of the ISTH- and IFE-specific transcription factors *Gata6* and *Ahr*[13] (Extended Data Fig. 6h), suggesting that Wnt has an early role in mediating the identity switch.

Control of cell identity is thought to be regulated through binding and remodelling of chromatin regions termed super-enhancers, which control the expression of *Lhx2* and *Klf5* in hair follicle stem cells[17]. Strikingly, vismodegib-sensitive ATAC peaks were found in super-enhancers at both loci (Fig. 3k). The *Lhx2* super-enhancer contains a GLI binding site that closed upon vismodegib treatment, whereas the *Klf5* super-enhancer harbours two TCF binding sites that became more accessible. Accordingly, the expression of *Lhx2* and *Klf5* was modulated by vismodegib treatment (Fig. 3l). Collectively, our data support an escape mechanism of drug-induced cell identity switching in BCC that features a broadly open chromatin structure that remains sensitive to the activity of key transcription factors.

We next set out to identify environmental cues that drive cell identity switching in BCC and, on the basis of the above results, focused on Wnt. Consistent with published reports showing that Wnt is required during BCC formation[18,19], initiating tumours expressed the universal Wnt target *Axin2*[20] (Fig. 4a). Unexpectedly, *Axin2* was suppressed in established tumours, but reactivated in residual disease (Fig. 4a). Levels of *Axin2* expression were inversely correlated with levels of *Wif1*, a potent secreted inhibitor of Wnt signalling[21] (Fig. 4a, b), and were induced as early as 2 days after the initiation of treatment (Fig. 4b), well before significant tumour shrinkage occurred (Extended Data Fig. 8a, b). Expression of *Wnt4* and *Wnt10a*, which have been implicated in the maintenance of normal basal epidermis[22], followed the same pattern (Fig. 4b). Human tumours showed a similar pattern; *AXIN2* mRNA was detectable in BCCs only after vismodegib treatment (Extended Data Fig. 8c–h). In addition to *Lhx2*, we also observed a return to pre-treatment levels of *Axin2* and *Wif1* in tumours that had been released from drug for 12 days (Extended Data Fig. 8i), again highlighting the plasticity of residual tumours.

We next tested whether treatment of BCC with a combination of vismodegib and a Wnt pathway inhibitor would have an effect on residual disease. We used a function-blocking anti-LRP6 antibody[23] (Extended Data Fig. 9a–c), as our RNA-seq analysis revealed that LRP6 was the predominant Wnt co-receptor in BCC. Treatment with vismodegib and anti-LRP6 led to a 33% decrease in the number of residual tumour nests compared to vismodegib monotherapy (Fig. 4c, d). Notably, the residual tumour burden correlated with the magnitude of Wnt pathway inhibition, as mice with fewer lesions experienced a greater decrease in *Axin2* expression (Extended Data Fig. 9d). In addition, Oil Red O staining was enhanced within residual tumours after combination treatment (Extended Data Fig. 9e–h), consistent with attenuation of Wnt signalling being required for sebocyte differentiation[24].

Finally, we investigated whether combined inhibition of Hh and Wnt signalling altered the regrowth of residual BCCs after treatment. Mice were dosed with BrdU to label tumour nests and were then treated

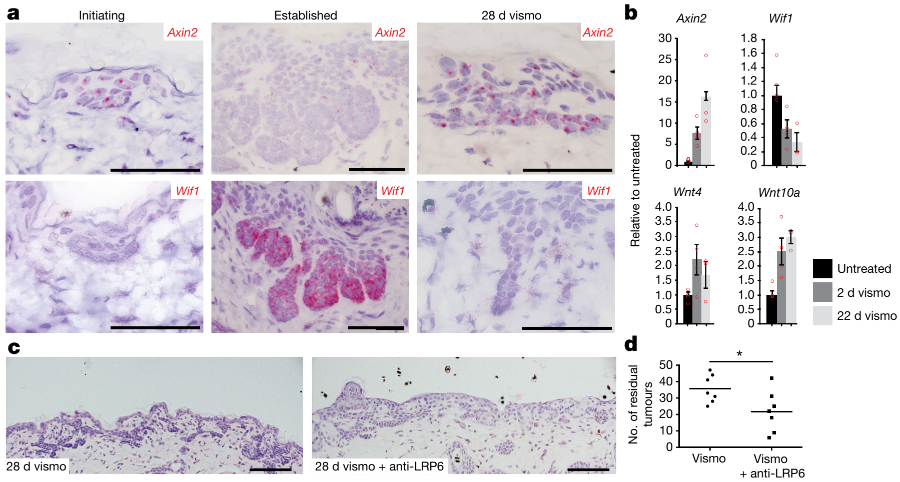


Fig. 4 | Wnt signalling is required for maintaining residual disease. **a**, Representative images of *Axin2* and *Wif1* ISH on mouse skin during initiating (3-week-old), established (8-week-old) and residual stages of BCC ($n = 4$ per stage). **b**, Relative expression levels (mean ± s.e.m.) of *Axin2*, *Wif1*, *Wnt4*, and *Wnt10a* in untreated BCCs and tumours treated for 2 or 22 days ($n = 3$ per group). **c**, Representative images of H&E-stained skin sections showing the amount of residual BCC after indicated treatments ($n = 7$ per group). **d**, Average residual tumour counts per length of skin in BCC mice from **c**. A two-tailed unpaired *t*-test was used to determine significance between groups ($*P = 0.0328$). All experiments were performed at least twice. Scale bars, 100 μm; *n* represents the number of mice.

for 28 days with either vismodegib alone or vismodegib and anti-LRP6. After discontinuation of treatment, tumour regrowth was delayed (but not abolished) in mice that were treated with combination therapy relative to those treated with vismodegib alone (Extended Data Fig. 9i–q), probably owing to incomplete Wnt inhibition (Extended Data Fig. 9d).

Together, our results show that despite the strong anti-tumour activity of vismodegib in BCC, residual tumour cells persist. Tumours display robust Hh pathway inhibition during treatment, indicating that the drug continues to block signalling. As such, residual tumour cells have not acquired drug resistance through de novo mutations, but have adopted an identity that no longer relies on Hh signalling. When treatment is discontinued, tumour cells reactivate the Hh pathway and resume growth. Although BCC patients experience clinical benefit from Hedgehog pathway inhibitors after a drug holiday[2], this approach may lead to the development of de novo resistance if additional mutations in the Hh pathway are acquired[25]. It is therefore important to achieve complete elimination of all residual tumour cells. The Wnt pathway seems to be critical for cell identity switching in BCC through reprograming of super-enhancers that drive the expression of key transcription factors. However, tolerability of complete Wnt pathway inhibition remains a challenge[26]. The ability of residual tumour cells to adopt a state that allows them to survive while remaining fully quiescent may represent a more widespread mechanism of resistance to targeted therapies[27]. Strategies that block this process may provide an opportunity to increase the rate of complete responses.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0596-y.

Received: 19 July 2017; Accepted: 22 August 2018;
Published online 8 October 2018.

1. Sekulic, A. et al. Efficacy and safety of vismodegib in advanced basal-cell carcinoma. *N. Engl. J. Med*. **366**, 2171–2179 (2012).
2. Sofen, H. et al. A phase II, multicenter, open-label, 3-cohort trial evaluating the efficacy and safety of vismodegib in operable basal cell carcinoma. *J. Am. Acad. Dermatol*. **73**, 99–105.e1 (2015).
3. Wang, G. Y., Wang, J., Mancianti, M. L. & Epstein, E. H. Jr. Basal cell carcinomas arise from hair follicle stem cells in Ptch1$^{+/-}$ mice. *Cancer Cell* **19**, 114–124 (2011).
4. Rubin, A. I., Chen, E. H. & Ratner, D. Basal-cell carcinoma. *N. Engl. J. Med*. **353**, 2262–2269 (2005).
5. Robarge, K. D. et al. GDC-0449-a potent inhibitor of the hedgehog pathway. *Bioorg. Med. Chem. Lett*. **19**, 5576–5581 (2009).
6. Von Hoff, D. D. et al. Inhibition of the hedgehog pathway in advanced basal-cell carcinoma. *N. Engl. J. Med*. **361**, 1164–1172 (2009).
7. Basset-Seguin, N. et al. Vismodegib in patients with advanced basal cell carcinoma (STEVIE): a pre-planned interim analysis of an international, open-label trial. *Lancet Oncol*. **16**, 729–736 (2015).
8. Vidal, V. P., Ortonne, N. & Schedl, A. SOX9 expression is a general marker of basal cell carcinoma and adnexal-related neoplasms. *J. Cutan. Pathol*. **35**, 373–379 (2008).
9. Feil, R., Wagner, J., Metzger, D. & Chambon, P. Regulation of Cre recombinase activity by mutated estrogen receptor ligand-binding domains. *Biochem. Biophys. Res. Commun*. **237**, 752–757 (1997).
10. Folgueras, A. R. et al. Architectural niche organization by LHX2 is linked to hair follicle stem cell function. *Cell Stem Cell* **13**, 314–327 (2013).
11. Nijhof, J. G. et al. The cell-surface marker MTS24 identifies a novel population of follicular keratinocytes with characteristics of progenitor cells. *Development* **133**, 3027–3037 (2006).
12. Pichery, M. et al. Endogenous IL-33 is highly expressed in mouse epithelial barrier tissues, lymphoid organs, brain, embryos, and inflamed tissues: *in situ* analysis using a novel Il-33-LacZ gene trap reporter strain. *J. Immunol*. **188**, 3488–3495 (2012).
13. Joost, S. et al. Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Syst*. **3**, 221–237.e9 (2016).
14. Ge, Y. et al. Stem cell lineage infidelity drives wound repair and cancer. *Cell* **169**, 636–650.e14 (2017).
15. Wang, S. et al. Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protocols* **8**, 2502–2515 (2013).
16. Kim, T. H. et al. Broadly permissive intestinal chromatin underlies lateral inhibition and cell plasticity. *Nature* **506**, 511–515 (2014).
17. Adam, R. C. et al. Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature* **521**, 366–370 (2015).
18. Yang, S. H. et al. Pathological responses to oncogenic Hedgehog signaling in skin are dependent on canonical Wnt/β3-catenin signaling. *Nat. Genet*. **40**, 1130–1135 (2008).
19. Youssef, K. K. et al. Adult interfollicular tumour-initiating cells are reprogrammed into an embryonic hair follicle progenitor-like fate during basal cell carcinoma initiation. *Nat. Cell Biol*. **14**, 1282–1294 (2012).
20. Jho, E. H. et al. Wnt/β-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway. *Mol. Cell. Biol*. **22**, 1172–1183 (2002).
21. Hsieh, J. C. et al. A new secreted protein that binds to Wnt proteins and inhibits their activities. *Nature* **398**, 431–436 (1999).
22. Lim, X. et al. Interfollicular epidermal stem cells self-renew via autocrine Wnt signaling. *Science* **342**, 1226–1230 (2013).
23. Tian, H. et al. Opposing activities of Notch and Wnt signaling regulate intestinal stem cells and gut homeostasis. *Cell Reports* **11**, 33–42 (2015).
24. Niemann, C. Differentiation of the sebaceous gland. *Dermatoendocrinol* **1**, 64–67 (2009).
25. Yauch, R. L. et al. Smoothened mutation confers resistance to a Hedgehog pathway inhibitor in medulloblastoma. *Science* **326**, 572–574 (2009).
26. Kahn, M. Can we safely target the WNT pathway? *Nat. Rev. Drug Discov*. **13**, 513–532 (2014).

27. Oser, M. G., Niederst, M. J., Sequist, L. V. & Engelman, J. A. Transformation from non-small-cell lung cancer to small-cell lung cancer: molecular drivers and cells of origin. *Lancet Oncol.* **16**, e165–e172 (2015).

Acknowledgements We thank J. Diaz for mouse colony management; the IVP team for animal dosing; S. Flanagan and V. Nunez for necropsy support; the Pathology core for help with histology; S. Biswas for help with LCM; the NGS laboratory for RNA-seq and ATAC–seq; I. Caro for help with human samples; and J. Svärd, R. Toftgard, and S. Teglund for *Ptch1*$^{fl/fl}$ mice.

Reviewer information *Nature* thanks R. Shivdasani and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.B., G.J.P.D., S.E.G. and F.J.d.S. conceptualized the project, designed experiments and analysed data. B.B. performed IHC, immunofluorescence and ISH on skin sections, isolated tumour nests by LCM and tumour cells by FACS, and performed qRT–PCR. G.J.P.D. performed initial characterization of residual disease. R.P. performed bioinformatics analyses. B.A. monitored mice for tumour burden and coordinated drug dosing. S.B. analysed ATAC–seq data. F.P. performed histopathological evaluation of skin sections. B.B., G.J.P.D. and F.J.d.S. wrote the manuscript.

Competing interests All authors are employees of Genentech and own shares of Roche.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0596-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0596-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to F.J.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

BCC mice and dosing. All mouse experiments were performed according to the animal use guidelines of Genentech, a Member of the Roche Group, conforming to California state legal and ethical practices. BCC mice were generated using the following alleles[28–30]: *K14*Cre^{ER}, *Ptch1*$^{fl/fl}$ (a generous gift from R. Toftgard and S. Teglund), and *p53*$^{fl/fl}$. BCC mice were singly housed and monitored for signs of advanced tumour burden, such as a scruffy coat and ear thickening. Tumour formation occurred in the absence of tamoxifen, probably owing to inherent leakiness of the Cre^{ER} recombinase[9]. Vismodegib was formulated as a suspension in 0.5% methyl-cellulose, 0.2% tween-80 (MCT). Minimal efficacious dosing of vismodegib by oral gavage was determined to be 75 mg/kg bodyweight twice a day, and this dosing schedule was used for all experiments. Anti-LRP6 bi-specific antibody[23] was given once daily at 30 mg/kg bodyweight by intraperitoneal (IP) injection. For label-retaining experiments, mice were given BrdU injections (100 μl of 10mg/ml) for three consecutive days before the start of treatment to label all dividing cells in the skin. Subsequent treatment with vismodegib stopped proliferation and prevented the dilution of incorporated BrdU in tumour nests. Lineage tracing with the lox-stop-lox TdTomato reporter allele (Jackson strain code: 007914) was initiated in mice at weanling age with three consecutive daily doses of tamoxifen (100 μl of 20 mg/ml in sunflower oil).

BCC histology score. A pathologist determined the extent of BCC tumour burden using the following scoring system: 0, minimal basaloid nests involve <10% of linear extent of skin; 1, isolated basaloid nests are generally confined to superficial half of dermis (do not extend to base of hair bulb); 2, basaloid nests crowd superficial dermis and/or extend into deeper half of dermis, but do not expand to fill dermal space completely in areas of deeper penetration; 3, basal cells pack dermis in confluent masses in >50% of skin.

RNA sequencing. Hundreds of lesions were laser capture micro-dissected (LCM) from multiple skin dissections per animal. Adjacent skin sections were stained for Ki67 and counterstained with haematoxylin to determine the proliferation status of tumour nests from untreated BCC mice. Three groups, each consisting of five mice, were included: 1) Ki67$^+$ tumours from untreated BCC mice; 2) Ki67$^-$ tumours from untreated BCC mice; and 3) residual tumours from BCC mice that were treated with vismodegib for 28 days. LCM samples were pooled on a per mouse basis and total RNA was extracted with the RNeasy kit (Qiagen). The concentration and integrity of the RNA were determined by NanoDrop 8000 (Thermo Scientific) and with a Bioanalyzer RNA 6000 Pico Kit (Agilent), respectively. cDNA was generated with the Ovation RNA-seq System V2 (Nugen) and sheared to 150–200 bp size using an LE220 focused ultrasonicator (Covaris). Fragment length was confirmed with the Bioanalyzer DNA 1000 Kit (Agilent) and samples were quantified with the Qubit dsDNA BR Assay (Life Technologies). A total of 1 μg of sheared cDNA and the TruSeq RNA Sample Preparation Kit v2 (Illumina) were used for the end repair step of each library. Library size was confirmed with High Sensitivity D1K screen tape on a 2200 TapeStation (Agilent Technologies), and concentration was determined by qPCR using the Library quantification kit (KAPA). Libraries were multiplexed and sequenced on a HiSeq 2500 (Illumina) to generate 30 million single-end 50-bp reads per library. Reads were first aligned to ribosomal RNA sequences to remove ribosomal reads. The remaining reads were aligned to the mouse reference genome (NCBI Build 38) using GSNAP[31] version '2013-10-10', allowing a maximum of two mismatches per 50 base pair sequence (parameters: '-M 2 -n 10 -B 2 -i 1 -N 1 -w 200000 -E 1–pairmax-rna = 200000–clip-overlap'). Transcript annotation was based on the RefSeq database (NCBI Annotation Release 104). To quantify gene expression, the number of reads mapped to the exons of each RefSeq gene was calculated. Read counts were scaled by library size, quantile normalized and precision weights calculated using the 'voom' R package[32]. Subsequently, differential expression analysis on the normalized count data was performed using the 'limma' R package[33] by contrasting vismodegib-treated samples with untreated samples. Gene expression levels were considered significantly different across groups if we observed $|\log_2[\text{fold change}]| \geq 1$ (estimated from the model coefficients) associated with a false discovery rate (FDR)-adjusted $P \leq 0.05$. Gene expression was obtained in form of nRPKM as described previously[34].

ATAC library preparation. ATAC–seq was performed as follows: 100,000 sorted cells were collected in 1 ml PBS + 3% FBS at 4 °C. Cells were centrifuged, then cell pellets were resuspended in 100 μl lysis buffer (Tris HCl 10 mM, NaCl 10 mM, MgCl2 3 mM, Igepal 0.1%) and centrifuged (500*g*) for 25 min at 4 °C. Supernatant was discarded and nuclei were resuspended in 50 μl reaction buffer (Tn5 transposase 2.5 μl, TD buffer 22.5 μl and 25 μl H_2O – Nextera DNA sample preparation kit, Illumina). The transposase reaction was performed for 30 min at 37 °C and then blocked by addition of 5 μl clean up buffer (NaCl 900 mM, EDTA 300 mM). DNA was purified using the MinElute purification kit (QIAGEN). DNA libraries were PCR amplified (Nextera DNA Sample Preparation Kit, Illumina), and size selected for 200 to 800 bp (BluePippin, Sage Sciences) following the manufacturers' protocols.

ATAC–seq. The libraries were sequenced on Illumina HiSeq 2500 sequencers. We obtained an average of 100 million paired-end reads (50 bp) per sample. Reads were aligned to the mouse reference genome (NCBI Build 38) using GSNAP[31] version '2013-10-10', allowing a maximum of two mismatches per read sequence (parameters: '-M 2 -n 10 -B 2 -i 1–pairmax-dna = 1000–terminal-threshold = 1000–gmap-mode = none–clip-overlap'). Reads aligning to locations in the mouse genome that contain substantial sequence homology to the MT chromosome or to blacklisted regions identified by the ENCODE consortium were omitted from downstream analyses. Properly paired reads derived from non-duplicate sequencing fragments were used to quantify chromatin accessibility according to the ENCODE pipeline standards with minor modifications as follows. Accessible genomic locations were identified by calling peaks with Macs2[35] using insertion-centred pseudo-fragments (73 bp; community standard) generated on the basis of the start positions of the mapped reads. Accessible peak locations were identified as described: in brief, we called peaks on a group-level pooled sample containing all pseudo-fragments observed in all samples within each group. Peaks in the pooled sample that were independently identified in two or more of the constituent biological replicates were retained for downstream analysis, using the union of all group-level reproducible peaks (https://www.encodeproject.org/atac-seq/#standards). We quantified the level of chromatin accessibility within each peak for each replicate as the number of pseudo-fragments that overlapped the peak in question and normalized these estimates using the TMM method[36]. To better understand the within- and between-group similarities in chromatin accessibility, we calculated the Pearson correlation coefficient for the top 5,000 most variable peaks and performed hierarchical clustering using the correlation measures.

We identified differentially accessible peaks between groups in the framework of a linear model implemented with the limma R package[33] and incorporating precision weights calculated with the voom function in the limma R package[32]. Peaks that showed an increase in accessibility at either day 6 or day 14 of vismodegib treatment were called vismodegib peaks and control peaks were called untreated peaks. For subsequent analysis, peaks were divided into promoter regions (1 kb up- and 2 kb down-stream of transcription start sites) and enhancers (peaks outside of promoter regions).

We identified enriched transcription factor (TF) motifs using HOMER v4.7[37]. To evaluate the significance of the TF enrichment we defined peaks as significantly differentially accessible based on a range of FDR adjusted *P* value thresholds between 1 and 0.01 and an $|\log_2[\text{fold change}]|$ in accessibility ≥ 1 (estimated from the model coefficients). Given the strong enrichment of the top motifs across a wide range of *P* value cutoffs, we decided to consider peaks as different across groups for a $|\log_2[\text{fold change}]| \geq 1$ and FDR adjusted *P* value ≤ 0.05 in subsequent analyses.

Chromatin accessibility was visualized as coverage tracks across genomic regions using the Integrated Genomics Viewer[38] and as heatmaps using DeepTools v3.0.1[39].

Regulatory potential (BETA analysis). We associated accessible chromatin regions with nearby genes using BETA[15]. The BETA minus mode was used to calculate the regulatory potential (determined through a distance-weighted measure) of specific sets of peaks within a certain distance to a target gene. The BETA basic mode allowed us to integrate differential expression with chromatin openness to evaluate whether the direct effect of changes in the chromatin landscape is promoting or repressing gene expression. In this mode all genes within 100 kb of a peak set are ranked (and listed along the *x*-axis) based on the regulatory potential using the ATAC–seq data. Subsequently, expression information is used to divide genes into downregulated by vismodegib (purple line), upregulated by vismodegib (red line) and transcriptionally unchanged (dashed line) genes. A one-tailed Kolmogorov–Smirnov test[40] was used to determine whether the upregulated and downregulated groups differed significantly from the group of transcriptionally unchanged genes.

FACS sorting of tumour cells. Back skin was taken from BCC mice harbouring the *Lgr5*^{GFPDTR} allele[41] and processed into a single-cell suspension[42]. Cells were stained for CD34 and the live/dead marker SYTOX blue. GFP-only tumour cells were sorted on a BD Biosciences Influx 'jet-in-air' cell sorter equipped with a combination of 355-, 405-, 488-, 561-, and 640-nm lasers. Fluorescence emission was collected through 460/50 (Sytox Blue), 530/40 (GFP), and 670/30 (CD34-Alexa647) bandpass filters. Sorting was performed with an 86-μm nozzle, a pressure of 30 psi and a frequency of 48,610 Hz. Setup and alignment of the instrument were performed using 3 μm ultra rainbow polystyrene beads (URFP-30-20, Spherotech Inc.). See Extended Data Fig. 6 for gating strategies.

Skin epithelial signatures. We downloaded microarray studies that profiled gene expression across various skin epithelial compartments, including GSE15185[43], GSE21568[44], GSE40612[19], GSE41704[45], GSE20269[46] and E-MTAB-1606[47]. Data normalization was performed using either the affy or lumi R packages, and differential expression analysis was performed using the limma R package[33].

Genes were called significantly differentially expressed when $|\log_2[\text{fold change}]| > 1$ and adjusted $P < 0.1$. Genes were considered bulge-specific if they were significantly upregulated in either the *Lgr5*$^+$ vs. *Lgr6*$^+$ comparison (GSE20269) or the CD34$^+$ vs. GFP$^-$ comparison (E-MTAB-1606). Genes were considered ISTH- or IFE-specific if they were significantly downregulated in the GSE20269 or E-MTAB-1606 studies, respectively. The ORS signature was obtained by downloading the gene count data from RNA-seq GEO study GSE90847[48] and identifying genes that were significantly upregulated ($\log_2[\text{fold change}] \geq 1$ and adjusted $P < 0.01$) in ORS as compared to hair follicle stem cells using the same voom and limma procedure as described above.

Gene set analysis. We performed QuSAGE[49] to identify relevant biological processes associated with vismodegib treatment. For that purpose we compared vismodegib-treated samples with untreated samples. For each comparison we then calculated the gene set activity (the mean difference in $\log_2$ expression of the individual genes that comprise the set) for the four sets of bulge, isthmus, IFE and ORS marker genes identified from the above public microarray experiments. In addition, we assessed the up- or downregulation of bulge, isthmus, IFE and ORS genes using gene set enrichment analysis (GSEA)[40]. The significance of the enrichment (shown as FDR) was determined through 1,000 permutations of random gene sets.

Immunohistochemistry and Immunofluorescence staining. Primary antibodies and their dilutions used in this study: rabbit Sox9 1:300 (Millipore, AB5535), rabbit Keratin1 1:1,000 (Covance, PRB-165 [AF109]), chicken Keratin5 1:2,000 (Biolegend, 90501), rabbit Keratin10 1:1,000 (Biolegend, 90541), rabbit Loricrin 1:1,000 (Covance, PRB-145P), goat IL-33 1:500 (R&D Systems, AF3626), rabbit Ki67 1:500 (GeneTex, GTX16667 [SP6]), rat BrdU 1:400 (BioRad, MCA2060), rabbit CC3 1:400 (Cell Signaling, 9661), and human Ep-CAM undiluted (Ber-EP4; Ventana 760-4383). Dorsal skin was shaved, dissected, and fixed in either 4% paraformaldehyde (PFA) or 10% neutral buffered formalin overnight. Fixed skin was washed in PBS and 70% ethanol, processed and embedded into paraffin. Fresh frozen skin was cut into strips and embedded in OCT (Sakura). Skin sections were cut at 6 μm on either a Leica RM2255 microtome or a Leica CM 3050 S cryostat. Fresh frozen skin sections were fixed in 4% PFA for 10 min before staining. For IHC, slides were de-paraffinized in xylenes, re-hydrated, and boiled in Dako target retrieval buffer for 10 min. Samples were then blocked with Dako protein-free blocking solution for 10 min and primary antibodies were diluted in Dako antibody diluent and exposed to samples overnight at 4 °C. Secondary antibodies (Invitrogen, Molecular Probes) for immunofluorescence staining were also diluted in Dako antibody diluent. IHC was performed using a Dako Envision$^+$ system-HRP polymer detection kit.

In situ hybridization. Traditional RNA ISH was performed on paraffin-embedded tissue sections using digoxigenin-labelled probes according to standard protocols[50]. Alkaline phosphatase activity was detected on tissue sections using BM Purple staining solution (Roche) after overnight incubation in alkaline phosphatase buffer.

Hybridizations using the RNAscope method were performed according to the manufacturer's protocol (Advanced Cell Diagnostics) using the RNAscope 2.5 HD Reagent Kit-RED (322350). Probes used were Mm*Lhx2* (485791), Mm*Axin2* (400331), Hu*AXIN2* (400241), Mm*DefB6* (430141), Mm*Wif1* (412361), and Mm*Gli1* (311001).

Taqman quantitative real-time PCR. Tumours were collected from fresh frozen skin sections by LCM. Total RNA was extracted using the RNeasy micro kit (Qiagen) and cDNA was prepared using the TaqMan RNA to Ct 1 Step kit (Applied Biosystems). TaqMan analysis was performed on an ABI7900HT (Applied Biosystems), data were analysed with SDS 2.3 software (Applied Biosystems) and normalized to *Hprt* transcript levels. The following Applied Biosystems Taqman assays were used: *Axin2* (Mm00443610_m1), *Wif1* (Mm00442355_m1), *Wnt4* (Mm01194003_m1), *Wnt10a* (Mm00839783_m1), *Lhx2* (Mm00839783_m1) and *Plet1* (MTS24 antigen; Mm01170995_m1).

Quantification of Oil Red O$^+$ and BrdU$^+$ tumours. The number of residual tumours in a linear unit of backskin equal to four lengths of a 5× magnification field of view was counted. Oil Red O$^+$ and BrdU$^+$ residual tumours were expressed as a fraction of the total number of tumours. BCCs were considered Oil Red O$^+$ when they contained at least one Oil Red O$^+$ cell and were not associated with a hair follicle (to avoid contamination from sebaceous glands). BCCs were considered BrdU$^+$ when they contained at least 20 intense BrdU$^+$ nuclei and did not contain any cells with a more weak, diluted nuclear signal (suggesting proliferation).

Statistics and reproducibility. See individual Methods sections for specific statistical methods. Experiments were independently repeated at least twice leading to similar results. No statistical method was used to predetermine sample size, and no mice were excluded from the analysis. BCC mice were age matched and randomly assigned to control and treatment groups. The investigators were blinded during outcome assessment. The data meet the assumptions of the statistical tests used and are presented as mean ± s.e.m.; $P \leq 0.05$ was considered statistically significant. For RNA-seq, differential expression analysis was performed using the 'limma' R package, which uses the moderated *t*-statistic for significance analysis. The adjusted *P* value was calculated using the Benjamini and Hochberg method to account for the false discovery rate.

Human subject data. The study (ClinicalTrials.gov identifier: NCT01201915) was conducted per FDA regulations, International Conference on Harmonization E6 Guideline for Good Clinical Practice, and applicable local, state, and federal laws. The protocol was approved by institutional review boards where applicable. Patients gave written informed consent. The study evaluated the activity of vismodegib in patients with smaller operable BCC by measuring the rate and durability of complete histologic clearance (CHC) of lesions. Patients with new, operable, nodular basal cell carcinoma received vismodegib (150 mg/d) followed by excision and Mohs micrographic surgery to ensure clear margins. Samples from cohort 1 were analysed for this study: patients received vismodegib for 12 weeks and target sites were immediately excised by standard means, followed by Mohs micrographic surgery to obtain clear margins. Biospecimen type: skin; anatomical site: scalp/head/neck or upper aspect of trunk, greater trunk; disease status of patients: new, nodular BCC, operable; clinical characteristics of patients: alive; clinical diagnosis of patients: nodular BCC by biopsy; pathology diagnosis: nodular BCC by biopsy; collection mechanism: excision, Mohs surgery; type of stabilization: formalin fixed; type of long-term preservation: formalin fixation; constitution of preservative: 10% neutral-buffered formalin.

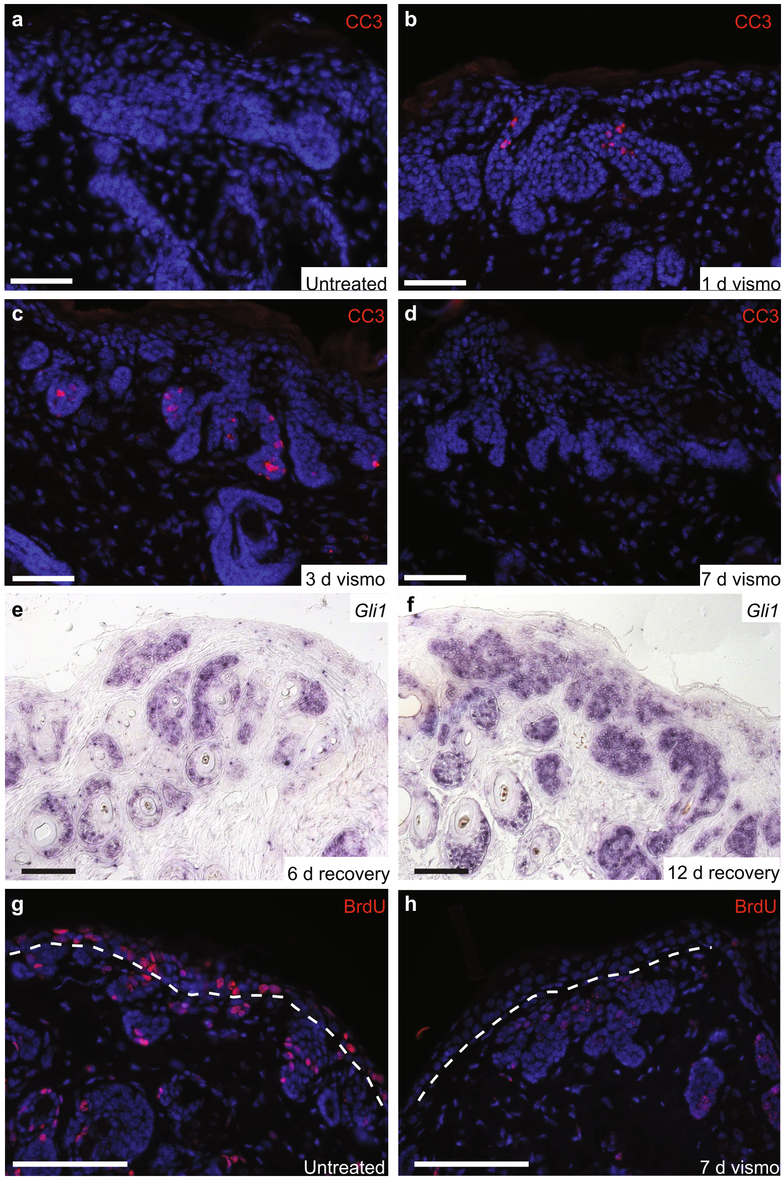
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

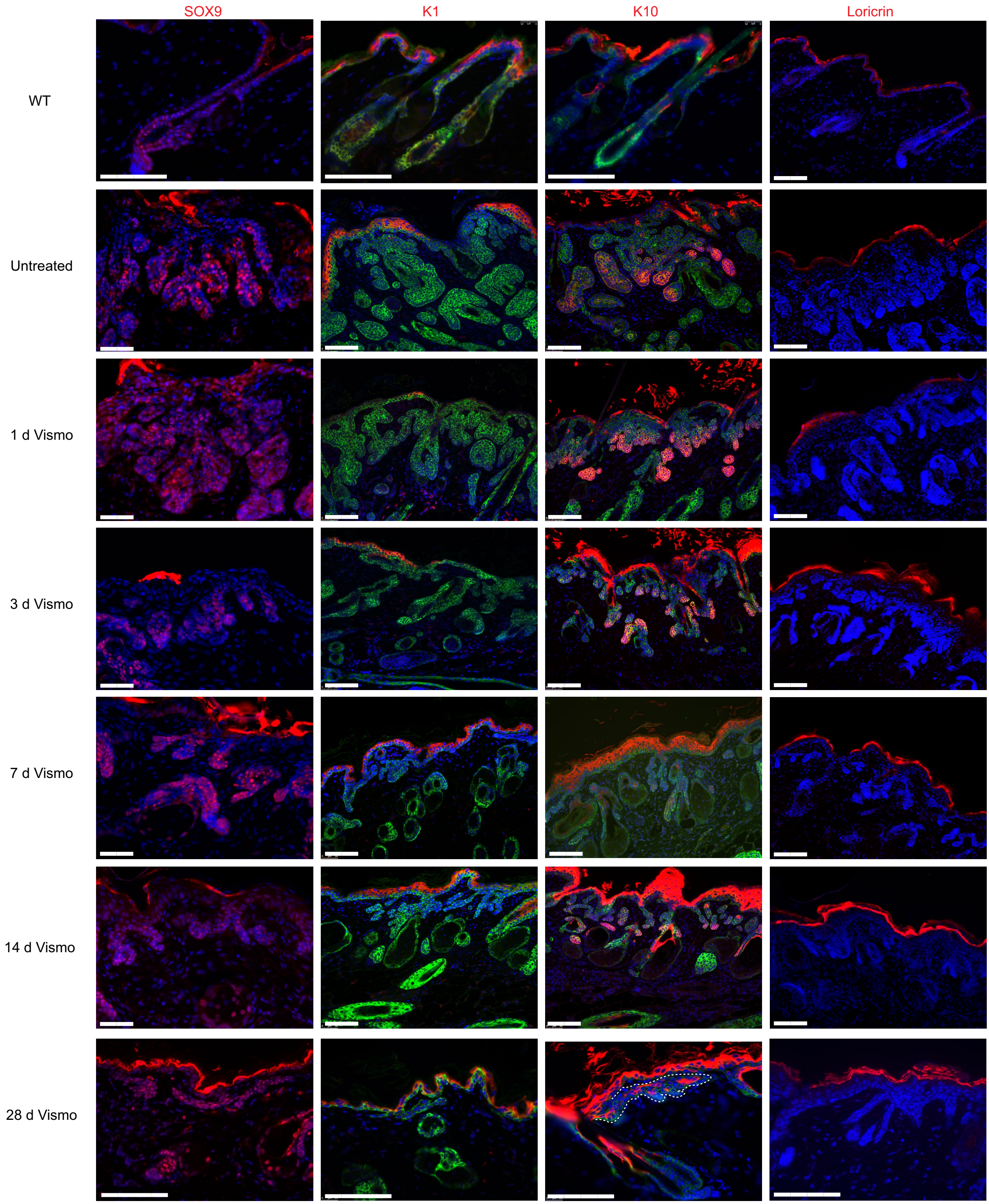
RNA-seq and ATAC–seq data supporting the findings of this study have been deposited in the Gene Expression Omnibus (GSE116966). All other data are available from the corresponding author upon reasonable request.

28. Vasioukhin, V., Degenstein, L., Wise, B. & Fuchs, E. The magical touch: genome targeting in epidermal stem cells induced by tamoxifen application to mouse skin. *Proc. Natl Acad. Sci. USA* **96**, 8551–8556 (1999).
29. Kasper, M. et al. Wounding enhances epidermal tumorigenesis by recruiting hair follicle keratinocytes. *Proc. Natl Acad. Sci. USA* **108**, 4099–4104 (2011).
30. Marino, S., Vooijs, M., van Der Gulden, H., Jonkers, J. & Berns, A. Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of Rb in the external granular layer cells of the cerebellum. *Genes Dev.* **14**, 994–1004 (2000).
31. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
32. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
33. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
34. Srinivasan, K. et al. Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. *Nat. Commun.* **7**, 11295 (2016).
35. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
36. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
37. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
38. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
39. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
40. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
41. Tian, H. et al. A reserve stem cell population in small intestine renders Lgr5-positive cells dispensable. *Nature* **478**, 255–259 (2011).
42. Jensen, K. B., Driskell, R. R. & Watt, F. M. Assaying proliferation and differentiation capacity of stem cells using disaggregated adult mouse epidermis. *Nat. Protocols* **5**, 898–911 (2010).
43. Greco, V. et al. A two-step mechanism for stem cell activation during hair regeneration. *Cell Stem Cell* **4**, 155–169 (2009).
44. Garza, L. A. et al. Bald scalp in men with androgenetic alopecia retains hair follicle stem cells but lacks CD200-rich and CD34-positive hair follicle progenitor cells. *J. Clin. Invest.* **121**, 613–622 (2011).
45. Blanpain, C., Lowry, W. E., Geoghegan, A., Polak, L. & Fuchs, E. Self-renewal, multipotency, and the existence of two cell populations within an epithelial stem cell niche. *Cell* **118**, 635–648 (2004).
46. Snippert, H. J. et al. Lgr6 marks stem cells in the hair follicle that generate all cell lineages of the skin. *Science* **327**, 1385–1389 (2010).
47. Page, M. E., Lombard, P., Ng, F., Göttgens, B. & Jensen, K. B. The epidermis comprises autonomous compartments maintained by distinct stem cell populations. *Cell Stem Cell* **13**, 471–482 (2013).

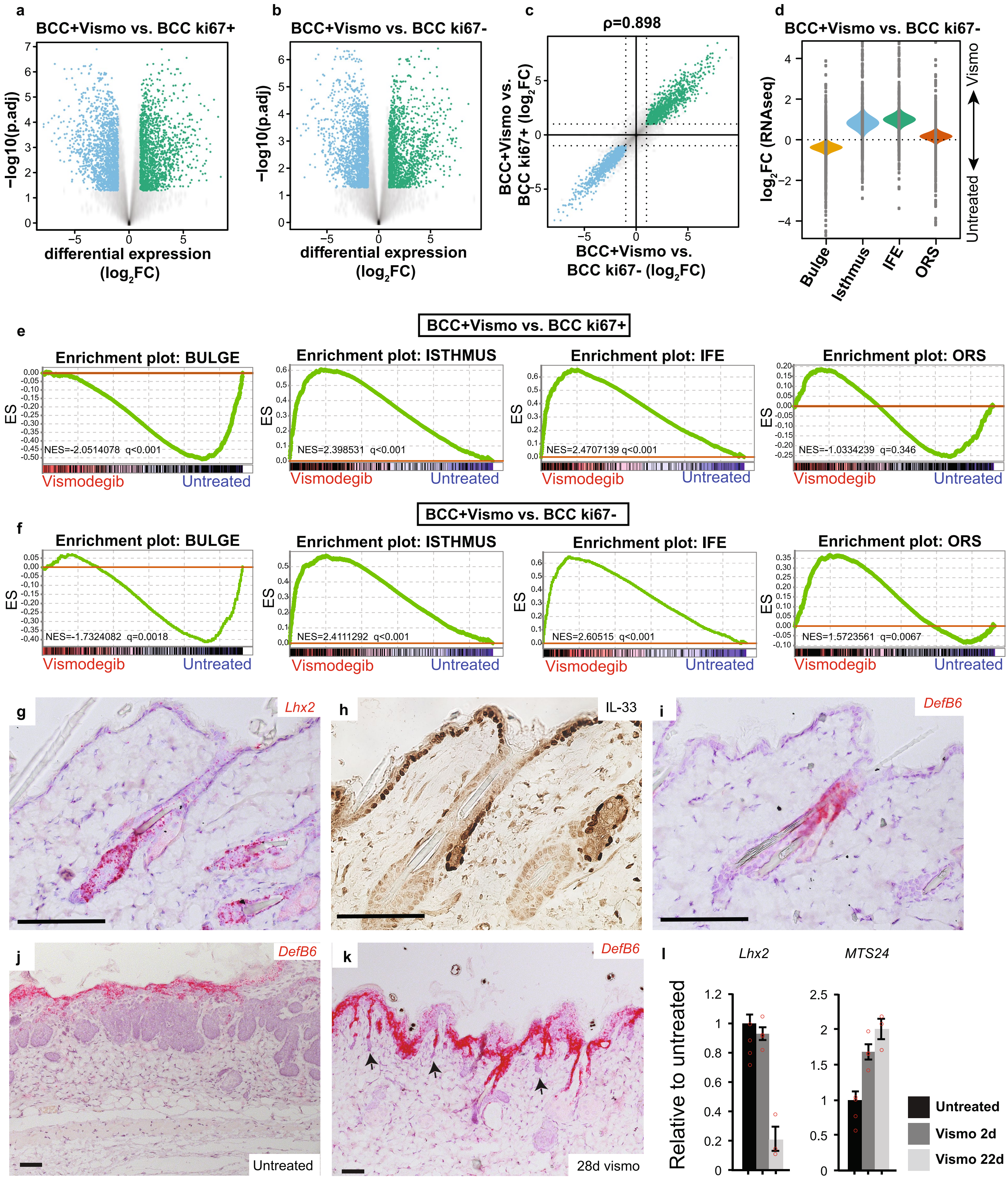
48. Yang, H., Adam, R. C., Ge, Y., Hua, Z. L. & Fuchs, E. Epithelial-mesenchymal micro-niches govern stem cell lineage choices. *Cell* **169**, 483–496.e13 (2017).
49. Yaari, G., Bolen, C. R., Thakar, J. & Kleinstein, S. H. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic Acids Res.* **41**, e170 (2013).
50. O'Neill, J. W. & Bier, E. Double-label in situ hybridization using biotin and digoxigenin-tagged RNA probes. *Biotechniques* **17**, 870, 874–875 (1994).
51. Beer, T. W., Shepherd, P. & Theaker, J. M. Ber EP4 and epithelial membrane antigen aid distinction of basal cell, squamous cell and basosquamous carcinomas of the skin. *Histopathology* **37**, 218–223 (2000).



Extended Data Fig. 1 | Limited apoptosis in tumour nests early during vismodegib treatment and recovery of residual BCCs. Related to Figs. 1a–d, 2e–h. **a**–**d**, Time-course analysis of cell death in our *K14*Cre^{ER};*Ptch1*$^{fl/fl}$;*Trp53*$^{fl/fl}$ mouse model of BCC during vismodegib treatment. Representative images of skin sections collected after indicated length of vismodegib treatment were stained for cleaved caspase 3 (CC3, red) and nuclei (DAPI, blue) ($n = 3$ per time point). The apoptosis marker CC3 is lacking in untreated BCC (**a**), but could be detected in rare tumour cells after just 1 day of treatment (**b**). CC3$^+$ tumour cells were most abundant after 3 days of treatment (**c**) but were no longer detectable after 7 days of treatment (**d**), indicating that additional mechanism(s) contribute to the tumour de-bulking that occurs at later time points. **e**, **f**, Additional images of *Gli1* ISH on skin sections from vismodegib-treated BCC mice shown in Fig. 2a. Note that the robust *Gli1* expression endures when animals remain off drug for 6 (**e**) or 12 days (**f**). **g**, **h**, Representative images of skin sections stained for BrdU (red) and nuclei (DAPI, blue) from untreated BCC mice and mice treated with vismodegib for 7 days ($n = 3$ per group). **g**, BrdU dosing before vismodegib treatment labels proliferating cells in BCCs and in the epidermis (above dotted line). **h**, Label-retaining BrdU$^+$ cells can be found only in residual tumour nests (below dotted line) after 7 days of vismodegib treatment, as epidermal cells continue to proliferate in the presence of drug and dilute out the label. All experiments were performed at least twice. Scale bars, 100 μm; n represents the number of mice.

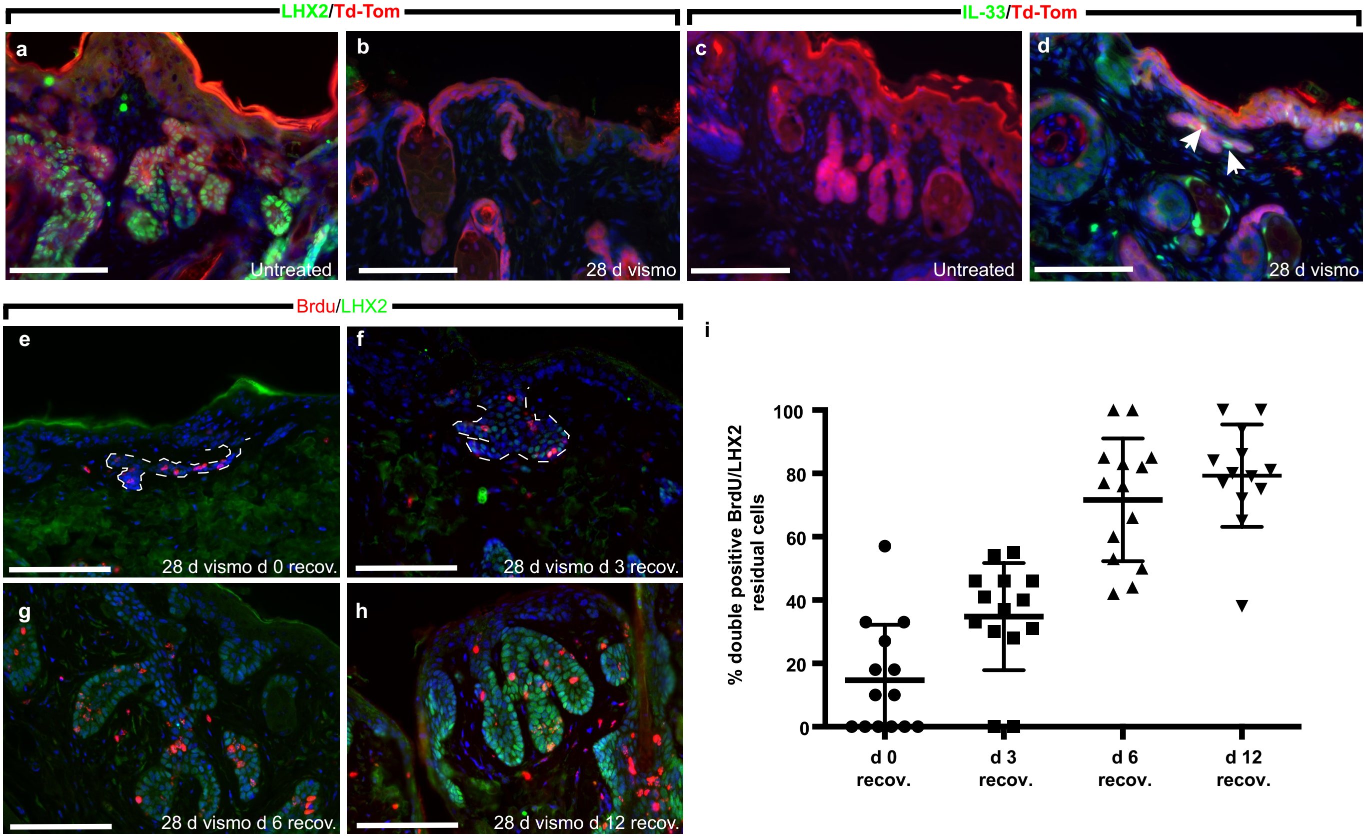


Extended Data Fig. 2 | Vismodegib treatment does not cause BCCs to differentiate into normal epidermis. Related to Fig. 1. Detailed time-course analysis of epidermal differentiation marker expression in BCC. Skin samples were collected from BCC mice after indicated length of vismodegib treatment. Representative images of skin sections are shown for $n = 4$ per time point. Residual tumour nests could be identified with the BCC marker SOX9 throughout treatment (column 1; red). By contrast, the supra-basal marker KRT1 (column 2; red) and the terminal differentiation marker loricrin (column 4; red) were at no point expressed in either untreated or residual BCCs. The supra-basal marker KRT10 was detected in both untreated and residual BCC (column 3; red). Its expression pattern within tumour nests gradually changed from a proximal/dermal location in untreated BCCs to the centre of residual BCCs, where it no longer overlapped with the basal marker KRT5 (green; outlined in 28 d vismo). Scale bars, 100 μm; n represents the number of mice. DAPI nuclear stain is blue.

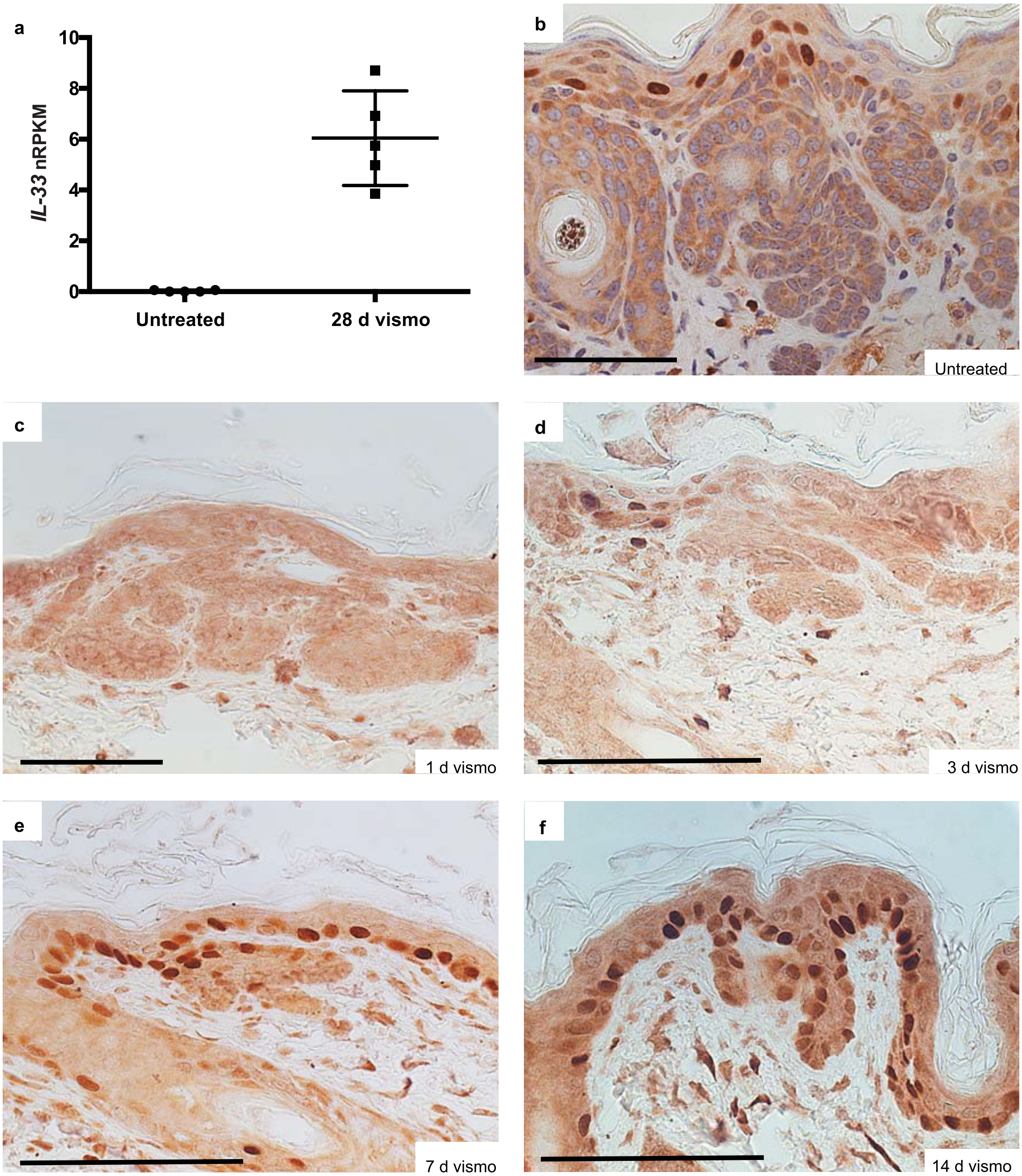


Extended Data Fig. 3 | See next page for caption.

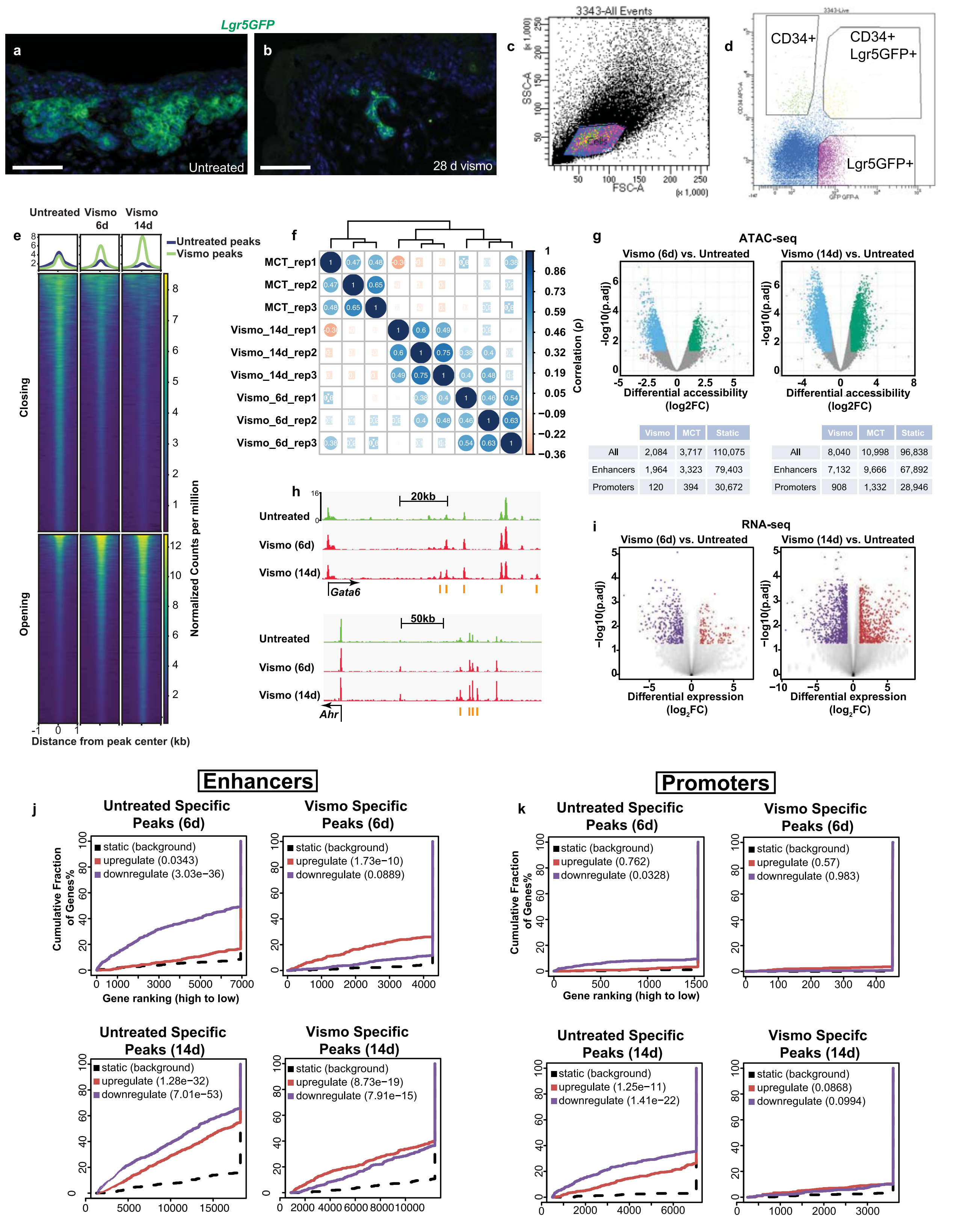
Extended Data Fig. 3 | Identification of differentially expressed genes in BCC after vismodegib treatment and their normal expression pattern in telogen skin. Related to Fig. 3. **a**, **b**, Volcano plots depicting the number of differentially expressed genes that are either significantly upregulated (green dots) or downregulated (blue dots) in residual disease after 28 days of vismodegib treatment relative to either Ki67$^+$ (**a**) or Ki67$^-$ (**b**) untreated BCC. Significance was tested with a moderated *t*-statistic (two-sided) and *P* values were adjusted for multiple testing with the Benjamini–Hochberg procedure ($\log_2$FC ≥ 1, adjusted $P \leq 0.05$; $n = 5$ per group). **c**, Robust correlation between the two differential gene expression estimates from the comparisons made in **a** and **b**. **d**, QuSAGE of indicated signatures in vismodegib-treated versus Ki67$^-$ untreated BCC samples ($n = 5$ per group). Coloured violins depict the differential expression of the entire gene set quantified by a probability density function calculated using QuSAGE. Each grey dot represents the $\log_2$[fold change] in expression of a particular gene in the signature. **e**, **f**, GSEA ranking of indicated skin compartment-specific gene signatures using the differential gene expression estimates from the comparisons made in **a**, **b**. Changes in expression were determined for $n = 5$ per group. **g**–**i**, Expression pattern of select skin epithelial markers in wild-type telogen skin. Representative images are shown for $n = 4$ per condition. **g**, *Lhx2* mRNA is particularly enriched in the bulge of hair follicles. **h**, Antibody stain revealing nuclear localization of IL-33 in the isthmus and interfollicular epidermis. **i**, *DefB6* mRNA is particularly enriched in the isthmus region of hair follicles. **j**, **k**, Representative images of *Defb6* ISH on skin from untreated BCC mice (**j**) and animals treated with vismodegib for 28 days (**k**; $n = 4$ per condition). Arrowheads indicate *DefB6*$^+$ residual BCCs. **l**, Relative expression of *Lhx2* (hair follicle bulge) and *MTS24* (ISTH) in BCCs treated with vismodegib for the indicated length of time. Data are plotted as mean $\pm$ s.e.m.; $n = 3$ per condition. Scale bars, 100 μm; *n* represents the number of mice.



Extended Data Fig. 4 | Residual BCCs reinitiate the hair follicle bulge program when vismodegib treatment is discontinued. Related to Fig. 3. **a**–**d**, Lineage tracing in BCC mice with a Cre-inducible TdTomato reporter allele after 28 days of vismodegib treatment ($n = 3$ per group). Untreated TdTomato^{+} BCCs display robust LHX2 antibody staining (**a**), but completely lack IL-33 antibody staining (**c**). TdTomato^{+} residual BCCs lack LHX2 antibody staining (**b**), but stain positive for IL-33 protein (**d**). **e**–**h**, Return of LHX2 expression in residual BCCs after vismodegib treatment. Representative images of skin sections stained for BrdU (red), nuclei (DAPI, blue) and LHX2 (green) from BCC mice treated with vismodegib for 28 days that were allowed to recover for the indicated times ($n = 3$ per group). **e**, BrdU and LHX2 co-staining is virtually absent in residual BCC (outlined) immediately after treatment. **f**, LHX2 staining is weak in BrdU-labelled tumour nests after 3 days of recovery. **g**, LHX2 staining is more prominent in residual tumours after 6 days of recovery. **h**, LHX2 staining is the most robust and shows maximum overlap with the BrdU label after 12 days of recovery. **i**, Proportion of BrdU–LHX2 double-positive tumour cells in skin sections from **e**–**h**. Graph shows mean ± s.d. Each data point represents the proportion of BrdU-labelled cells that were LHX2-positive within a single tumour nest ($n \geq 12$ tumours from 3 mice per time-point). Scale bars, 100 μm; n represents the number of mice except for **i**, where it represents the number of tumours.



Extended Data Fig. 5 | *Il-33* expression in BCCs during vismodegib treatment. Related to Fig. 3. **a**, Expression of *Il33* in LCM tumours from untreated BCC mice and mice treated with vismodegib for 28 days ($n = 5$ per group). RNA-seq expression data were quantile normalized and represented as nRPKM. Data plotted are mean ± s.e.m.; data points (dots, squares) indicate *Il33* expression in tumours from individual mice. **b**–**f**, Detailed time-course analysis of IL-33 expression in BCCs during vismodegib treatment. Representative images of skin samples collected from both untreated and treated BCC mice are shown ($n = 3$ per treatment and time point). IL-33 gradually appears in residual BCCs over the course of treatment. This nuclear factor can readily be detected in the basal layer of the epidermis, but is lacking in untreated BCCs (**b**) and in tumours treated with vismodegib for just 1 (**c**) or 3 (**d**) days. Nuclear IL-33 begins to emerge in a fraction of the tumour nests after 7 days of treatment (**e**), and most residual BCCs express IL-33 by 14 days of treatment (**f**). Scale bars, 100 μm; *n* represents the number of mice.



Vismo (6d) vs. Untreated:

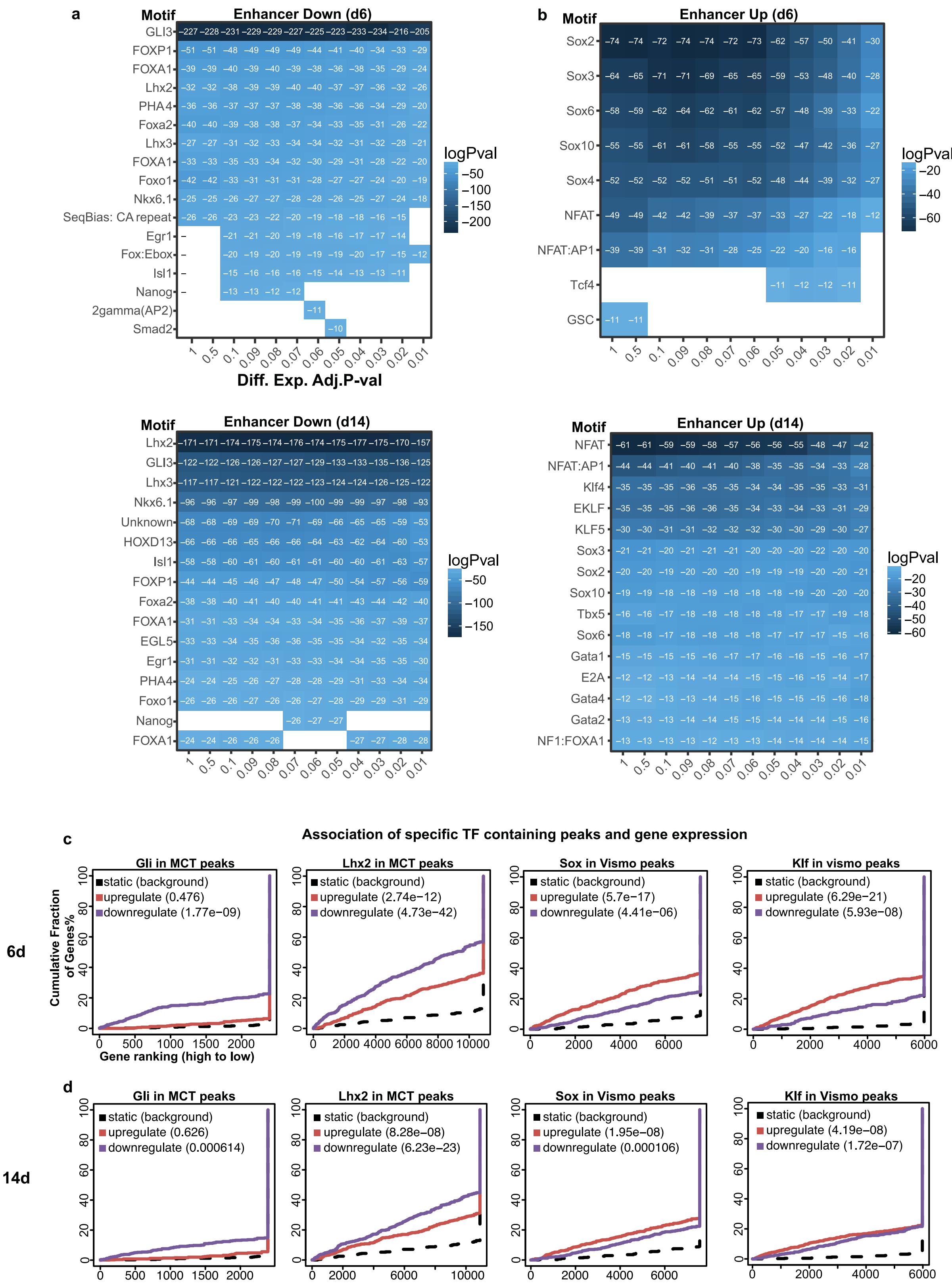
	Vismo	MCT	Static
All	2,084	3,717	110,075
Enhancers	1,964	3,323	79,403
Promoters	120	394	30,672

Vismo (14d) vs. Untreated:

	Vismo	MCT	Static
All	8,040	10,998	96,838
Enhancers	7,132	9,666	67,892
Promoters	908	1,332	28,946

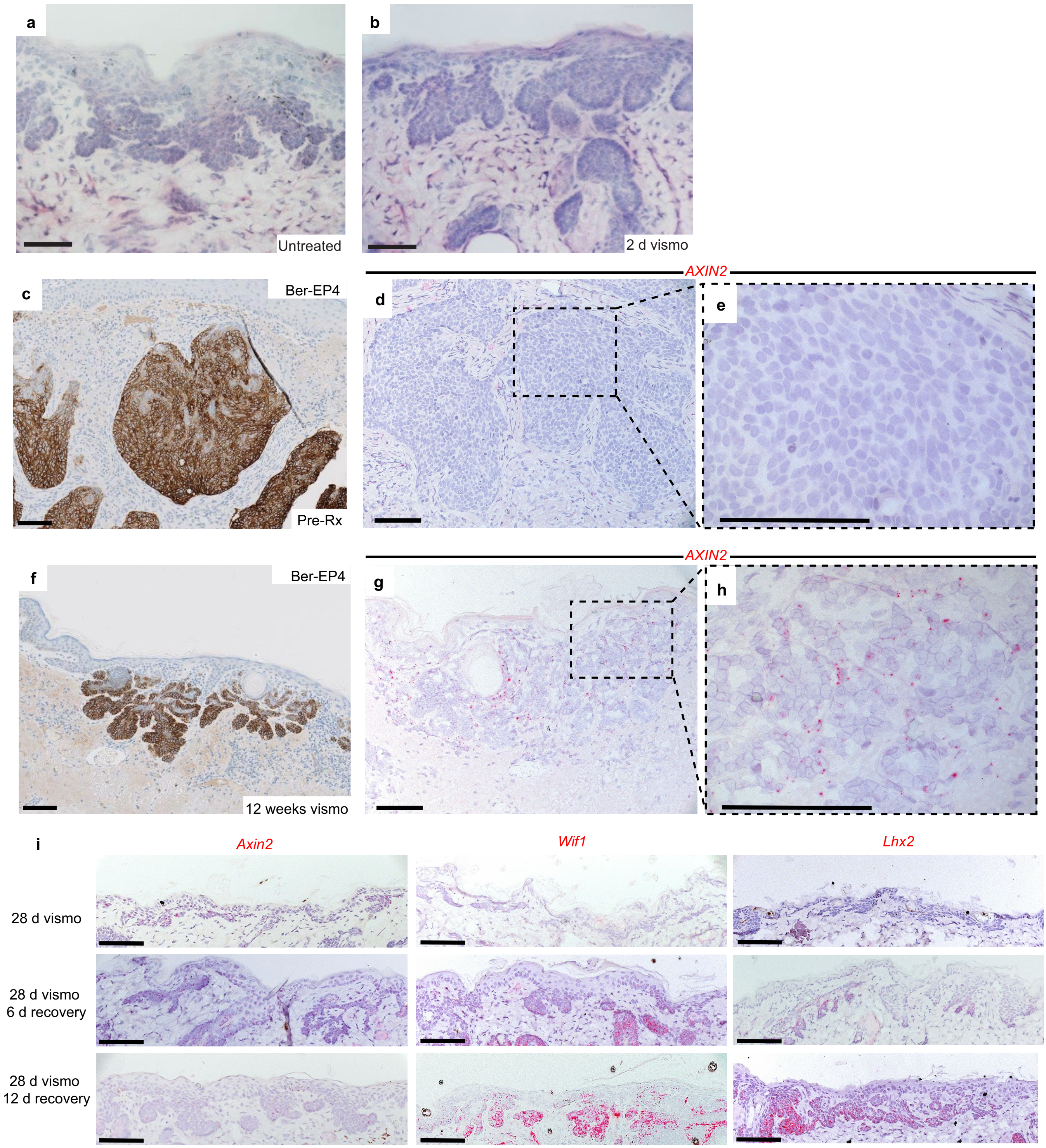
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | ATAC–seq of FACS-sorted BCC cells and regulatory potential of accessible chromatin regions. Related to Fig. 3. **a**, **b**, We crossed our mouse model of BCC to the *Lgr5*^{DTRGFP} allele[41], to enable FACS isolation of both untreated (**a**) and residual (**b**) tumour cells. Representative images of skin sections stained for GFP (green) and nuclei (DAPI, blue) are shown for $n = 3$ per group. **c**, Gating for low side scatter enriched for epidermal cells after dissociation of back skin samples[42]. **d**, Final gating strategy used for the FACS isolation of BCC cells. Lgr5^{+} hair follicle bulge stem cells were excluded by staining for CD34. **e**, Heat-map of differentially accessible chromatin regions that either close (top) or open (bottom) in response to indicated length of vismodegib treatment. Data are shown as normalized peak counts per million genomic DNA fragments averaged from $n = 3$ per condition. **f**, Matrix of Pearson correlation coefficients showing the overlap in chromatin accessibility between samples based on the 5,000 most variable peak regions. Pearson's ρ statistic was used to determine each correlation. **g**, Volcano plots depicting chromatin accessibility in sorted BCC cells after 6 (left) and 14 days (right) of vismodegib treatment relative to untreated (MCT) controls ($n = 3$ per group). Significance was tested with a moderated t-statistic (two-sided) and P values were adjusted for multiple testing with the Benjamini–Hochberg procedure. Each dot represents a peak (an open chromatin region). Common peaks are grey, and specific peaks are either blue (untreated) or green (6 and 14 days vismodegib). The table summarizes the number of open chromatin regions according to promoters and enhancers. **h**, Chromatin traces showing averaged ATAC peaks at the *Gata6* and *Ahr* loci; $n = 3$ per group. Orange rectangles mark TCF binding elements in chromatin that are sensitive to vismodegib. **i**, Volcano plots depicting the number of differentially expressed genes that are either significantly upregulated (red dots; 150 at 6 d and 794 at 14 d) or downregulated (purple dots; 416 at 6 d and 1,129 at 14 d) in sorted BCC cells after 6 (left) and 14 days (right) of vismodegib treatment relative to untreated controls. Significance was tested with a moderated t-statistic (two-sided) and P values were adjusted for multiple testing with the Benjamini–Hochberg procedure ($\log_2 FC \geq 1$, adjusted $P \leq 0.05$; $n = 3$ per group). **j**, **k**, BETA analysis graphs depicting the effect of differentially open enhancers (**j**) and promoter peaks (**k**) on gene expression in sorted BCC cells after 6 (top) and 14 days (bottom) of vismodegib treatment relative to untreated controls ($n = 3$ per group). Genes were ranked from high to low according to the regulatory potential of the corresponding chromatin peak. Purple lines represent vismodegib-downregulated genes, while red lines represent vismodegib-upregulated genes. A one-tailed Kolmogorov–Smirnov test[40] was used to determine whether the up- and downregulated groups differed significantly (shown as P values in parentheses) from the static group of transcriptionally unchanged genes (dashed lines). Scale bars, 100 μm; n represents the number of mice.

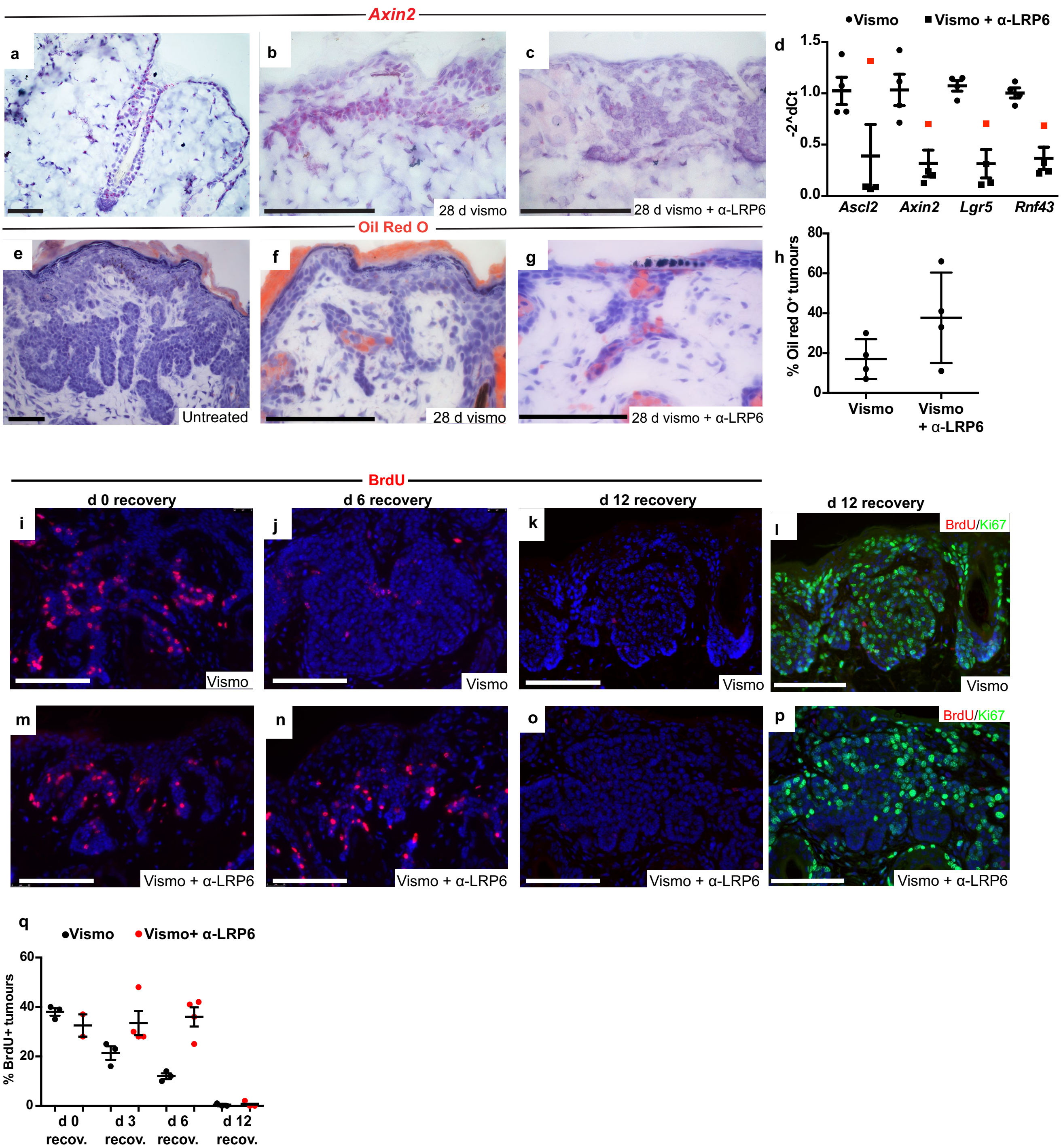


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Enrichment of transcription factor binding motifs in differentially regulated enhancers and their regulatory potential. Related to Fig. 3. **a**, **b**, Enrichment of transcription factor binding sites in non-promoter peaks that are either differentially closing (**a**) or opening (**b**) after 6 (top) and 14 days (bottom) of vismodegib treatment relative to untreated BCC cells ($n = 3$ per group). A hypergeometric test implemented in HOMER was used to identify enriched motifs, which are ranked by P value (top to bottom) over increasingly stringent P value cutoffs for peak calling (left to right). **c**, **d**, BETA analysis graphs depicting the effect of differentially open enhancer peaks with indicated transcription factor binding motifs on gene expression in sorted BCC cells after 6 (**c**) and 14 days (**d**) of vismodegib treatment relative to untreated controls ($n = 3$ per group). Genes were ranked from high to low according to the regulatory potential of the corresponding chromatin peak. Purple lines represent vismodegib-downregulated genes, while red lines represent vismodegib-upregulated genes. A one-tailed Kolmogorov–Smirnov test[40] was used to determine whether the up- and downregulated groups differed significantly (shown as P values in parentheses) from the static group of transcriptionally unchanged genes (dashed lines). n represents the number of mice.



Extended Data Fig. 8 | *AXIN2* is also induced in human residual BCC. Related to Fig. 4. **a**, **b**, Representative images of H&E-stained skin sections from untreated BCC mice (**a**) and from animals treated with vismodegib for 2 days (**b**; $n = 3$ per group). Note that the relatively short treatment has no impact on the overall morphology and size of the tumour nests. **c**–**h**, Representative images of Ber-EP4 (brown) staining and *AXIN2* ISH (red) on skin sections collected from a patient with BCC before and after 12 weeks of vismodegib treatment. The human BCC marker Ber-EP4[51] was used to identify nodular BCC in a screening biopsy (**c**) and residual BCC at the tumour site after treatment (**f**). **d**, **e**, **g**, **h**, Representative images of *AXIN2* ISH on skin sections immediately adjacent to the ones shown in **c** and **f**. **d**, Untreated nodular BCC lacks *AXIN2* mRNA. **g**, Residual BCC contains elevated *AXIN2* mRNA levels. **e**, **h**, High-magnification views of boxed regions in **d** and **g**, respectively. **i**, *Axin2*, *Wif1* and *Lhx2* ISH on skin sections from BCC mice after 28 days of vismodegib treatment. Skin samples were collected 0 (top), 6 (middle) and 12 (bottom) days after the end of treatment, and representative images are shown for $n = 3$ per group. Note that *Axin2* expression rapidly declines during the recovery period, whereas the expression of both *Wif1* and *Lhx2* progressively increase. All experiments were replicated at least twice. Scale bars: **c**, **f**, 90 μm; other panels, 100 μm.



Extended Data Fig. 9 | Effects of the vismodegib and anti-LRP6 combination treatment on mouse BCC. Related to Fig. 4. **a**–**c**, Representative images of *Axin2* ISH on skin sections from wild-type telogen mice (**a**) and from BCC mice treated for 28 days with either vismodegib alone (**b**) or vismodegib and anti-LRP6 (**c**). Note that co-treatment with anti-LRP6 strongly reduces *Axin2* mRNA levels in residual BCCs ($n = 3$ per group). **d**, Relative expression of Wnt target genes in the small intestine of BCC mice treated for 28 days with either vismodegib alone (circles) or vismodegib and anti-LRP6 (squares; $n = 4$ per group). Data are plotted as mean ± s.e.m. Three out of four BCC mice that received vismodegib and anti-LRP6 experienced at least an 80% reduction in expression of all four Wnt target genes that was accompanied by a low residual tumour burden (black squares; 6, 9 and 18 residual tumours per length of skin), while the one animal with poor Wnt pathway modulation had a much higher residual tumour burden (red square; 31 residual tumours per length of skin). **e**–**g**, Representative images of Oil Red O (orange)-stained skin sections from untreated BCC mice (**e**) and mice treated for 28 days with either vismodegib alone (**f**) or vismodegib and anti-LRP6 (**g**) (nuclear blue counter stain, haematoxylin; $n = 4$ per treatment). **h**, Proportion of Oil Red O$^+$ tumours per length of skin in BCC mice from **f**, **g**. Average percentages ± s.d. are shown for $n = 4$ per group. **i**–**p**, BrdU labelling of residual BCCs and their subsequent growth after cessation of indicated treatment. Representative images of residual BCCs stained for BrdU (red), nuclei (DAPI, blue) and Ki67 (green) are shown ($n \geq 2$ per treatment and time point) from BCC mice treated for 28 days with either vismodegib alone (**i**–**l**) or vismodegib and anti-LRP6 (**m**–**p**). Note that residual BCCs from both treatment groups are fully proliferating by 12 days of recovery and have diluted out the BrdU label. **q**, Proportion of BrdU$^+$ tumours per length of skin from BCC mice from **i**–**p**. Average percentages ± s.d. are shown for $n \geq 2$ per group. All experiments were replicated at least twice. Scale bars, 100 μm; n represents the number of mice.

Extended Data Table 1 | Skin epithelial signatures used in this study

Experiment	Comparison	# Sig. genes	Overlap p-value	Reference
GSE15185	DP vs. Bu	2399	0.5655	Greco et al., 2009
GSE15185	HG vs. Bu	1228	0.7609	Greco et al., 2009
GSE15185	DP vs. HG	2613	0.7872	Greco et al., 2009
GSE20269	ISTH vs. Bu	963	3.91×10^{-15}	Snippert et al., 2010
GSE21568	IFE vs. Bu	2668	9.06×10^{-51}	Garza et al., 2011
GSE40612	4wk BCC vs IFE	486	1.10×10^{-56}	Youssef et al., 2012
GSE40612	10wk BCC vs. IFE	111	1.07×10^{-4}	Youssef et al., 2012
GSE41704	Non-Bu epithelium vs. Bu	266	1.38×10^{-7}	Blanpain et al., 2004
MTAB1606	ISTH vs. Bu	426	3.31×10^{-19}	Page et al., 2013
MTAB1606	IFE vs. Bu	683	7.52×10^{-17}	Page et al., 2013
MTAB1606	ISTH vs. IFE	236	5.75×10^{-16}	Page et al., 2013

List of skin epithelial gene signatures with experiment IDs used in this study for comparison of differentially expressed genes from untreated and residual BCC. The number of genes within each signature is shown together with the *P* value associated with the overlap of the differentially expressed genes identified in this study. DP, dermal papillae; HG, hair germ.

LETTER

https://doi.org/10.1038/s41586-018-0603-3

A slow-cycling LGR5 tumour population mediates basal cell carcinoma relapse after therapy

Adriana Sánchez-Danés[1], Jean-Christophe Larsimont[1], Mélanie Liagre[1], Eva Muñoz-Couselo[2,3], Gaëlle Lapouge[1], Audrey Brisebarre[1], Christine Dubois[1], Mariano Suppa[4], Vijayakumar Sukumaran[1], Véronique del Marmol[4], Josep Tabernero[2,3] & Cédric Blanpain[1,5]*

Basal cell carcinoma (BCC) is the most frequent cancer in humans and results from constitutive activation of the Hedgehog pathway[1]. Several Smoothened inhibitors are used to treat Hedgehog-mediated malignancies, including BCC and medulloblastoma[2]. Vismodegib, a Smoothened inhibitor, leads to BCC shrinkage in the majority of patients with BCC[3], but the mechanism by which it mediates BCC regression is unknown. Here we used two genetically engineered mouse models of BCC[4] to investigate the mechanisms by which inhibition of Smoothened mediates tumour regression. We found that vismodegib mediates BCC regression by inhibiting a hair follicle-like fate and promoting the differentiation of tumour cells. However, a small population of tumour cells persists and is responsible for tumour relapse following treatment discontinuation, mimicking the situation found in humans[5]. In both mouse and human BCC, this persisting, slow-cycling tumour population expresses LGR5 and is characterized by active Wnt signalling. Combining *Lgr5* lineage ablation or inhibition of Wnt signalling with vismodegib treatment leads to eradication of BCC. Our results show that vismodegib induces tumour regression by promoting tumour differentiation, and demonstrates that the synergy between Wnt and Smoothened inhibitors is a clinically relevant strategy for overcoming tumour relapse in BCC.

Vismodegib (GDC0449) is the first Smoothened inhibitor to be approved for the treatment of locally advanced and metastatic BCC. A small fraction of patients does not respond to vismodegib administration: their tumours continue to grow and do not show inhibition of the Hedgehog (Hh) signalling pathway during vismodegib treatment[3]. This type of vismodegib resistance is frequently associated with genetic mutations that render vismodegib unable to inhibit the Hh pathway[6,7]. Most patients treated with vismodegib experience clinical benefits[3]. However, many patients respond only partially: their tumours initially regress under therapy but relapse after vismodegib discontinuation[3,5]. The mechanisms by which vismodegib induces tumour regression and that underlie non-genetic resistance to vismodegib therapy are unknown.

To study the mechanisms by which vismodegib leads to BCC regression, we induced BCC in mice by deleting *Ptch1* or overexpressing the constitutive active form of *Smo* (*SmoM2*) in the epidermis using *Krt14-CreER*[8,9]. BCCs induced by conditional knockout of *Ptch1* (*Ptch1*^{cKO}) arise mainly from the upper hair follicle (infundibulum) whereas those induced by *SmoM2* originate from the interfollicular epidermis (IFE)[4,8]. Eight weeks after deletion of *Ptch1* by tamoxifen administration, mice showing fully formed BCCs were treated daily with vismodegib and analysed at different time points (Fig. 1a). A decrease in tumour burden was observed during the first 5 weeks of vismodegib treatment, followed by stabilization of tumour size from 5 to 12 weeks, together with the appearance of vismodegib-persistent lesions (Fig. 1b, c, Extended Data Fig. 1a–d). Vismodegib administration led to the conversion of the BCCs into pre-neoplastic lesions (hyperplasia and dysplasia), which persisted as drug-tolerant lesions (Fig. 1d, Extended Data Fig. 1e). These results show that vismodegib induces tumour shrinkage and the progressive appearance of drug-tolerant lesions.

Staining for active caspase-3 two weeks after vismodegib administration showed a similar number of apoptotic cells in treated and untreated mice (Fig. 1e, f, Extended Data Fig. 1f, g), indicating that apoptosis is not the main mechanism by which vismodegib induces BCC regression. As quiescence has been described as a mechanism of cancer resistance to therapy[10], we assessed the proportion of Ki67-positive tumour cells and observed a strong decrease in the proportion of proliferative cells in persistent lesions (Fig. 1g, h, Extended Data Fig. 1h, i), suggesting that quiescence contributes to the emergence of drug-tolerant cells.

Lgr5 is expressed by different epithelial stem cells, including hair follicle stem cells (HFSCs)[11], and is upregulated during BCC initiation[9] (Extended Data Fig. 2a). In situ hybridization (ISH) showed that *Lgr5* was highly expressed in untreated BCCs and its expression persisted, albeit at a lower level, in vismodegib-tolerant lesions (Fig. 2a, Extended Data Fig. 2b)

ISH for *Gli1*, a transcription factor that relays Hh signalling and a Hh target gene, demonstrated that *Gli1* was co-expressed with *Lgr5* before treatment and was strongly downregulated in all tumour cells upon vismodegib treatment (Fig. 2a–c, Extended Data Fig. 2b–d), consistent with the strong inhibition of Hh signalling by vismodegib. Drug-tolerant lesions did not present mutations in *Smo*, the most frequently mutated gene in vismodegib-resistant BCC[6,7] (Extended Data Fig. 2e), reinforcing the notion that the persistence of drug-tolerant lesions is not mediated by mutations that abrogate vismodegib sensitivity, as it occurs in vismodegib-resistant BCCs that continue to grow during treatment[6,7].

Relapse of BCC upon vismodegib discontinuation has been reported in human patients[5]. Discontinuation of vismodegib administration for 4 weeks in *Krt14*^{CreER};*Ptch1*^{cKO};*Lgr5*$^{DTR-GFP}$ mice[12] bearing drug-persistent lesions led to the re-growth of BCCs to their pre-treatment size. Moreover, re-administration of vismodegib to mice with relapsing BCCs led to tumour regression (Fig. 2d).

To determine whether the quiescent tumour cell population mediates tumour relapse, we performed BrdU pulse-chase label retention studies by administrating BrdU for 3 days in mice with BCC to label proliferative cells, and then monitored the labelling during 5 weeks of vismodegib treatment. We found BrdU label-retaining cells (LRCs) in LGR5^{+} drug-tolerant lesions, suggesting that persisting tumour cells existed before vismodegib treatment and underwent a phenotype switch from a proliferative to a quiescent state (Fig. 2e, f). Upon discontinuation of vismodegib, relapsed tumours lost the LRCs (Fig. 2e, f), suggesting that quiescent LRCs actively proliferated, diluting the BrdU. To test this possibility directly, we performed BrdU–EdU double-labelling studies. Administration of EdU during vismodegib discontinuation led to EdU incorporation in the majority of the LGR5^{+}BrdU^{+} LRCs (Fig. 2g,

[1]Laboratory of Stem Cells and Cancer, Université Libre de Bruxelles, Brussels, Belgium. [2]Vall d'Hebron University Hospital, Universitat Autònoma de Barcelona, Barcelona, Spain. [3]Vall d'Hebron Institute of Oncology (VHIO), Universitat Autònoma de Barcelona, Barcelona, Spain. [4]Department of Dermatology, Erasme Hospital, Université Libre de Bruxelles, Brussels, Belgium. [5]WELBIO, Université Libre de Bruxelles, Brussels, Belgium. *e-mail: cedric.blanpain@ulb.ac.be

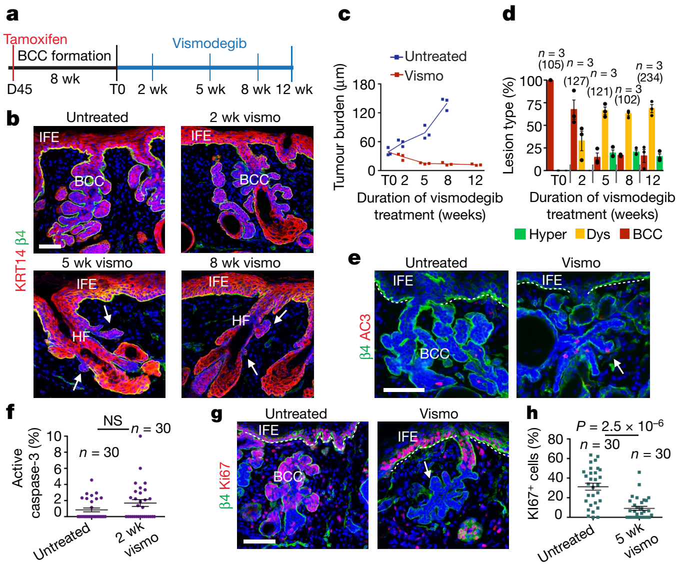


Fig. 1 | Slow-cycling tumour cells persist following vismodegib treatment in mouse *Ptch1*^{cKO}-derived BCCs. a, Protocol for tumour induction and vismodegib (vismo) administration. **b**, Immunostaining for KRT14 and β4-integrin (β4) in ventral skin from *Ptch1*^{cKO} mice. HF, hair follicle. **c**, Tumour burden (total area occupied by tumours divided by the length of the analysed epidermis) in untreated and vismodegib-treated mice ($n = 3$ mice analysed per time point and condition). Squares show data for individual mice, lines show mean. See Source Data. **d**, Quantification of lesion type (mean ± s.e.m.) upon vismodegib treatment ($n = 3$ mice, total number of lesions analysed per time point indicated in parentheses). Hyper, hyperplasia; dys, dysplasia. **e**, Immunostaining for active caspase-3 (AC3) and β4-integrin. **f**, Percentage of AC3^{+} tumour cells (mean ± s.e.m.) in untreated and vismodegib-treated mice ($n = 30$ lesions analysed from 3 mice). Two-sided *t*-test. **g**, Immunostaining for Ki67 and β4-integrin. **h**, Percentage of Ki67^{+} tumour cells (mean ± s.e.m.) in untreated and vismodegib-treated mice ($n = 30$ lesions analysed from 3 mice). Two-sided *t*-test. Hoechst nuclear staining in blue; scale bars, 50 µm. Dashed line in **e**, **g** delineates basal lamina. Arrows in **b**, **e**, **g** indicate vismodegib-persistent lesions.

Extended Data Fig. 2f, g), further demonstrating that the quiescent LRCs re-enter cell cycle and proliferate to contribute to tumour relapse.

To determine whether quiescence promotes the persistence of the vismodegib-tolerant lesions, we assessed whether increased epidermal proliferation decreased the number of drug-tolerant lesions. Mice bearing LGR5^{+} persistent lesions were treated for 2 weeks with vismodegib in combination with 12-*O*-tetradecanoylphorbol-13-acetate (TPA) or retinoic acid, two drugs that promote epidermal proliferation. Combined administration of vismodegib and TPA or retinoic acid promoted proliferation, which led to the elimination of LGR5^{+} persistent lesions (Extended Data Fig. 2h–j), demonstrating that when persistent slow-cycling cells are forced to proliferate they become sensitive to vismodegib and are eliminated.

We isolated the persistent tumour cells using fluorescence-activated cell sorting (FACS), by combining LGR5–GFP with LRIG1, which does not co-localize with LGR5 in resting hair follicles[13] (Extended Data Fig. 2k–m). Upon vismodegib administration, the proportion of LGR5^{+}LRIG1^{+} cells decreased and there was an increase in the LRG5^{-}LRIG1^{+} population (Extended Data Fig. 2m, n).

We then characterized the gene signature of FACS-isolated LGR5^{+}LRIG1^{+} and LRG5^{-}LRIG1^{+} tumour cell populations from untreated BCCs using microarray analysis. It has been shown that, during BCC initiation, IFE and infundibulum cells targeted by *Ptch1*^{cKO} or *SmoM2* are reprogrammed into fates resembling those of embryonic hair follicle progenitor (EHFP) cells and adult hair follicles in a Wnt-dependent manner[9,14]. Genes that were upregulated in LGR5^{+}LRIG1^{+} tumour cells compared to LRG5^{-}LRIG1^{+} tumour cells (LGR5^{+} BCC signature) overlapped significantly with the EHFP signature[15] (23.3%), resting HFSC signature[16] (16.4%) and LGR5^{+} hair follicle signature[17] (44.2%) (Fig. 3a, Extended Data Fig. 3a). The LGR5^{+} BCC signature included genes downstream of the Hh signalling pathway, such as *Ptch1*, *Ptch2* and *Hhip*, genes involved in the Wnt signalling pathway, such as *Lgr5*, *Fzd2* and *Lef1*, transcription factors expressed by EHFPs, such as *Runx1* and *Lhx2*, and genes expressed by HFSCs, such as *Tbx1* and *Foxc1* (Extended Data Fig. 2b). Immunostaining for LEF1, LHX2, CUX1, TBX1, and ALCAM in *Ptch1*^{cKO}-induced BCCs confirmed the increased expression of these Wnt signalling, EHFP and HFSC markers in LGR5^{+} tumour cells (Extended Data Fig. 3c).

To assess whether the LRG5^{-}LRIG1^{+} population represents a differentiated part of the BCC, we defined genes that were upregulated in LRG5^{-}LRIG1^{+} tumour cells compared to LGR5^{+}LRIG1^{+} tumour cells (LGR5^{-} signature). Notably, the LGR5^{-} signature overlapped significantly with previously reported LRIG1[13] and IFE[16] signatures, including markers of IFE or infundibulum differentiation such as *Ovol1*, *Notch3*, *Defb6*, *Krt1* and *Krt10* (Extended Data Fig. 3d, e). PCR analysis performed on FACS-isolated LGR5^{+}LRIG1^{+} and LGR5^{-}LRIG1^{+} tumour cells confirmed that both populations had *Ptch1* deletion, and staining for the proliferation marker Ki67 showed that the LRG5^{+}LRIG1^{+} population was more proliferative than the LGR5^{-}LRIG1^{+} population (Extended Data Fig. 3f, g).

To directly assess whether LGR5^{-}LRIG1^{+} cells were more differentiated than LGR5^{+}LRIG1^{+} cells, we performed transplantation assays of FACS-isolated tumour cell populations from *Krt14*^{CreER};*Ptch1*^{cKO};*Lgr5*$^{DTR-GFP}$ and *Krt14*^{CreER};*Ptch1*^{cKO};*Trp53*^{cKO};*Lgr5*$^{DTR-GFP}$ mice, which grow faster and form bigger tumours[18]. Groups of cells resembling early BCC and expressing KRT14, LGR5 and LRIG1 were observed only upon transplantation of LGR5^{+}LRIG1^{+} cells from *Trp53*^{cKO} mice (in three out of seven mice). By contrast, no tumour cells were observed following the transplantation of LGR5^{-}LRIG1^{+} cells from *Trp53*^{cKO} BCCs or in the absence of *Trp53* deletion (Extended Data Fig. 4a, b). Tumours found after transplantation of LGR5^{+}LRIG1^{+} cells mimicked the different cell types present in BCCs: LGR5^{+}LRIG1^{+}, LGR5^{-}LRIG1^{+} and cells with a flat differentiated morphology expressing keratin-10 (KRT10) (Extended Data Fig. 4b, c). Together, these results show that BCCs contain a more stem-like or progenitor-like tumour cell population (LGR5^{+}LRIG1^{+}) and a more differentiated population (LGR5^{-}LRIG1^{+}) of tumour cells. Immunostaining for the primary cilia marker ARL13B and the coactivator MKL1 showed that neither loss of primary cilia[19] nor serum response factor (SRF)–MKL1 activation[20] is involved in the drug-tolerant phenotype described here (Extended Data Fig. 5a–d).

To define the molecular mechanisms by which vismodegib promotes tumour shrinkage and appearance of drug-tolerant lesions, we compared the transcriptional profiles of FACS-isolated LGR5^{+}LRIG1^{+} and LGR5^{+}LRIG1^{-} tumour cells from untreated BCCs and mice that received vismodegib for 8 weeks. We found that the overlap between the LGR5^{+}LRIG1^{+} signature and the EHFP[15], LGR5^{+} hair follicle[17] and resting HFSC[16] signatures was considerably lower in vismodegib-treated cells than in untreated cells (Fig. 3a, b). Vismodegib treatment induced a strong decrease in the expression of Hh target genes such as *Gli1*, *Gli2*, *Ptch1*, *Ptch2* and *Hhip* (Fig. 3c). Only a small part of the reduction in overlap between the vismodegib-treated and EHFP signatures was driven by Hh target genes such as *Hhip1*, *Ptch2* and *Gli1*, and the reduction in overlap between the HFSC and vismodegib-treated signatures was not mediated by Hh target genes as the HFSC signature was obtained in the resting state, when Hh signalling is not active[16]. Genes found in the EHFP and HFSC signatures, such as *Runx1*, *Lhx2*, *Lgr5*, *Alcam* and *Tbx1* were also downregulated following vismodegib administration at the mRNA and protein levels (Fig. 3c and Extended Data Fig. 6a).

The overlap between the LGR5^{+}LRIG1^{+} signature and the infundibulum[13] and IFE[16] signatures increased considerably upon vismodegib treatment, with genes such as *Ovol1*, *Notch3*, *Plet1*, *Defb1*, *Defb6*, *Krt1* and *Krt10* being strongly upregulated after vismodegib treatment (Fig. 3d–f, Extended Data Fig. 6b), indicating that vismodegib

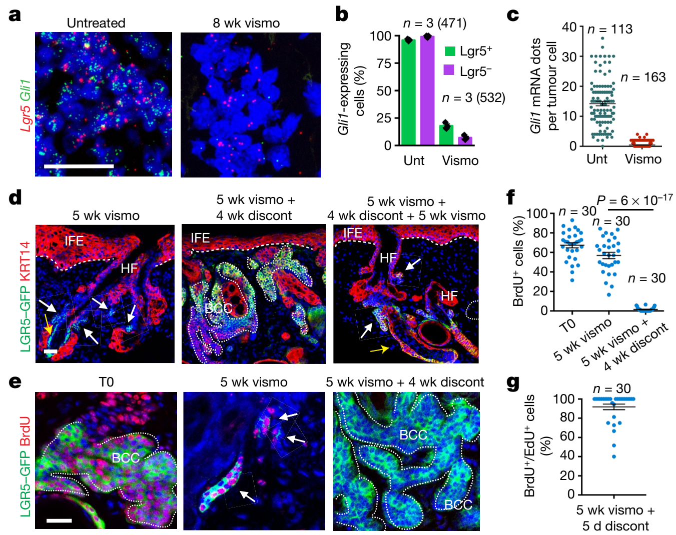


Fig. 2 | Slow-cycling LGR5⁺ LRCs mediate tumour relapse following discontinuation of vismodegib. a, ISH for *Lgr5* (red) and *Gli1* (green) in untreated and treated tumour cells from $Krt14^{CreER}$;$Ptch1^{cKO}$ mice. **b**, Percentage of tumour cells (LGR5⁺ and LGR5⁻) that express *Gli1* (mean ± s.e.m.; $n=3$ $Krt14^{CreER}$;$Ptch1^{cKO}$;$Lgr5^{DTR\text{-}GFP}$ mice, total number of cells analysed indicated in parentheses). **c**, Distribution of the number of *Gli1* mRNA dots per tumour cell with and without treatment (unt) (mean ± s.e.m.; $n=113$ and 163 total tumour cells from 3 $Krt14^{CreER}$;$Ptch1^{cKO}$;$Lgr5^{DTR\text{-}GFP}$ mice per condition and time point). **d**, Immunostaining for LGR5–GFP and KRT14 in $Ptch1^{cKO}$ ventral skin following vismodegib treatment, discontinuation and vismodegib re-administration. Three independent experiments per condition were analysed and showed similar results. **e**, Immunostaining for LGR5–GFP and BrdU following BrdU administration and after 5 and 9 weeks of chase in $Ptch1^{cKO}$-induced BCCs. **f**, **g**, Proportions of LGR5⁺ tumour cells presenting BrdU labelling at T0, after vismodegib treatment and discontinuation (**f**) and BrdU⁺EdU⁺ double-positive tumour cells 5 days after vismodegib discontinuation (**g**) in $Ptch1^{cKO}$-induced BCCs (mean ± s.e.m.; $n=30$ lesions analysed from 3 mice per condition). Two-sided *t*-test (**f**, **g**). Hoechst nuclear staining in blue; scale bars, 25 μm. Dashed lines delineate basal lamina; white arrows indicate vismodegib-persistent lesions; yellow arrows indicate hair follicle LGR5⁺ cells. RA, retinoic acid.

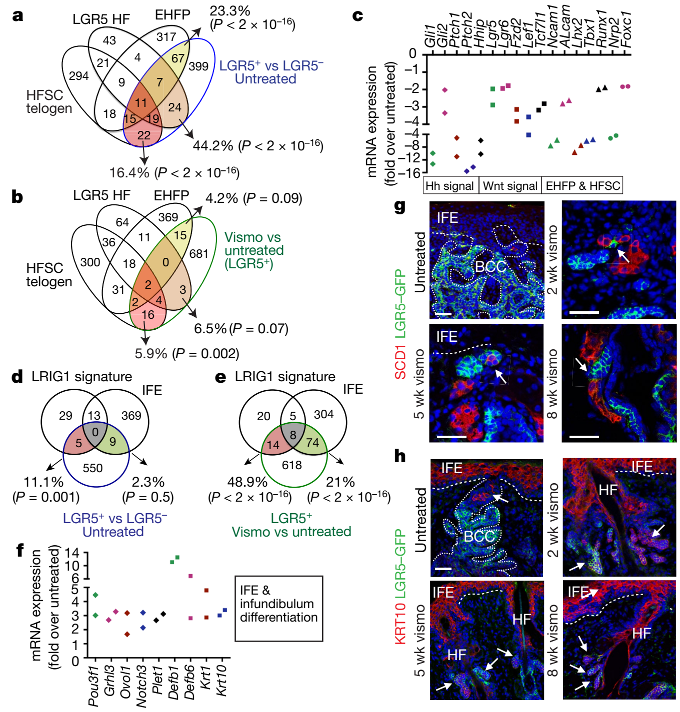


Fig. 3 | Vismodegib promotes BCC differentiation. a, **b**, Venn diagrams showing the similarities and the differences from two independent microarray experiments between genes upregulated more than twofold in LGR5⁺LRIG1⁺ versus LGR5⁻LRIG1⁺ cells (**a**) or in LGR5⁺LRIG1⁺ cells treated with vismodegib versus untreated (**b**) with the telogen HFSC signature[16], hair follicle LGR5-expressing cell signature[17] and EHFP signature[15]. **c**, mRNA expression of hair follicle genes downregulated in LGR5⁺LRIG1⁺ cells after 8 weeks of vismodegib administration ($n=2$ independent microarray experiments). **d**, **e**, Venn diagrams showing the similarities and differences between genes that were differentially upregulated more than twofold from two independent microarray experiments in untreated LGR5⁺LRIG1⁺ versus LGR5⁻LRIG1⁺ cells (**d**) or in untreated versus vimodegib-treated LGR5⁺LRIG1⁺ tumour cells (**e**) compared to IFE[16] and LRIG1[13] signatures. **f**, mRNA expression of IFE and infundibulum genes that were upregulated in LGR5⁺LRIG1⁺ cells after 8 weeks of vismodegib administration ($n=2$ independent microarray experiments). **g**, Immunostaining for LGR5–GFP and SCD1 in untreated and vismodegib-treated $Ptch1^{cKO}$-induced BCCs. Arrow indicates areas of sebaceous gland differentiation. **h**, Immunostaining for LGR5–GFP and KRT10 in untreated and vismodegib-treated $Ptch1^{cKO}$ mice. Arrow indicates differentiation of LGR5⁺ tumour cells into KRT10-expressing cells. Three independent experiments per condition were analysed with similar results (**g**, **h**). Hoechst nuclear staining in blue; scale bars, 50 μm. Dashed line delineates basal lamina. *P* values calculated using hypergeometric test for each intersection of two subsets of genes with phyper function in R software (**a**, **b**, **d**, **e**).

promotes the differentiation of BCC into IFE- and infundibulum-like cells, possibly through a Notch-dependent mechanism[21].

LRIG1⁺ stem cells give rise to infundibulum and sebaceous gland under homeostatic conditions[13]. We performed staining for sebaceous gland markers (SCD1 and adipophilin) and lipids (Oil Red O). Whereas sebaceous cysts were visible in the dermis under untreated conditions, cells expressing sebaceous gland markers were localized within the tumour mass after two weeks of vismodegib treatment and adjacent to the neoplastic lesions after five or eight weeks of treatment (Fig. 3g, Extended Data Fig. 6c, d). We studied the expression of KRT10 and *Defensin-β6* (*Defb6*), which are normally expressed in infundibulum and IFE cells. Upon vismodegib administration, KRT10 and *Defb6* were strongly upregulated in tumour cells (Fig. 3h, Extended Data Fig. 6e), consistent with vismodegib inducing tumour differentiation towards a sebaceous gland/infundibulum/IFE-like fate in $Ptch1^{cKO}$-derived BCCs.

We then assessed whether vismodegib also promotes differentiation of BCC into IFE in *SmoM2*-induced BCC. Upon vismodegib administration, *SmoM2*-expressing cells connected to normal differentiating IFE cells expressed high levels of the IFE differentiation marker keratin-1 (KRT1) (Extended Data Fig. 6f). We studied the effect of vismodegib administration on the survival and morphology of the *SmoM2* clones during BCC initiation. Two weeks after *SmoM2* expression, mice were treated daily with vismodegib for six weeks (Extended Data Fig. 7a). Vismodegib administration led to a progressive loss of *SmoM2*-expressing clones in comparison to untreated conditions and to the emergence of clones with normal differentiation, with only a small proportion of the clones progressing into hyperplasia and dysplasia (Extended Data Fig. 7b–d). The normally differentiated clones observed during vismodegib treatment were positive for the differentiation marker KRT10 but did not express LHX2, an HFSC marker that is found in hyperplasias and dysplasias (Extended Data Fig. 7e, f), indicating that vismodegib administration inhibits oncogene-induced hair follicle reprograming, promotes differentiation of *SmoM2*-expressing cells into an IFE-like fate and prevents BCC initiation.

To assess whether LGR5⁺ tumour cells consist of heterogeneous populations in terms of proliferation and differentiation, we isolated LGR5⁺LRIG1⁺ tumour cells on the basis of expression of the proliferation marker CD71[10] two weeks after vismodegib administration, when both persistent cells and cells that are responsive to vismodegib co-exist. The CD71⁺ population expressed higher levels of proliferation (*Ki67* and *Aurka*) and differentiation markers (*Krt1*, *Krt10* and

Scd1) (Extended Data Fig. 7g), indicating that the more proliferative tumour cells are more prone to vismodegib-induced differentiation. Immunostaining for the differentiation marker KRT10 in LGR5$^+$ tumour cells after BrdU label-retention followed by two weeks of vismodegib administration showed that the majority of BrdU-labelled cells were negative for KRT10, whereas KRT10 was observed in non-LRCs or in LRCs in which the BrdU signal was lower owing to its dilution following cell division (Extended Data Fig. 7h). These results support the notion that vismodegib induces a higher rate of differentiation in the drug-responsive tumour population that actively cycles.

To determine the relevance of our findings to human patients, we analysed biopsies from four patients with locally advanced BCCs before, during or immediately after discontinuation of vismodegib treatment. Vismodegib did not eradicate all tumour cells in these patients, and small tumorigenic lesions expressing *LGR5* persisted despite the administration of vismodegib for months (Extended Data Fig. 8a–c). ISH for *GLI1* and quantification of *GLI1* mRNA dots per tumour cell before, after or during vismodegib treatment showed that there was almost no *GLI1* expression in samples from patients during vismodegib treatment but few more *GLI1*-expressing cells were found shortly after discontinuation of vismodegib treatment (Extended Data Fig. 8c, d), indicating that vismodegib administration efficiently inhibits Hh signalling in these drug-persistent lesions. Ki67 immunohistochemistry showed that vismodegib-persistent lesions were more quiescent than untreated BCC cells, and vismodegib induced the expression of the differentiation marker KRT10 in human tumour cells (Extended Data Fig. 8e, f). Notably, patients 1 and 2 relapsed 6 and 9 months after treatment discontinuation, respectively, and patient 4 had previously relapsed after vismodegib discontinuation, showing that vismodegib-mediated tumour cell persistence is fully reversible upon drug withdrawal and re-inducible upon a new cycle of vismodegib treatment (Extended Data Fig. 8a). Together, these results show that drug-tolerant lesions exist in human BCC, characterized by the expression of LGR5 and relative quiescence.

To assess whether LGR5$^+$ cells mediate tumour growth, we lineage-ablated LGR5$^+$ tumour cells by administrating diphtheria toxin for 10 days to *Krt14*^{CreER};*Ptch1*^{cKO};*Lgr5*$^{DTR-GFP}$ mice and for 15 days to *Krt14*^{CreER};*Rosa*^{SmoM2};*Lgr5*$^{DTR-GFP}$ mice (Extended Data Fig. 9a). Diptheria toxin treatment could not be extended because LGR5 deletion is toxic to normal liver cells[12]. Diptheria toxin administration led to a substantial elimination of the tumour mass in both BCC models (80% of the initial tumour mass) and to almost total elimination of LGR5-expressing cells in *Ptch1*^{cKO}-induced BCC (Extended Data Fig. 9b–g), further demonstrating the importance of LGR5$^+$ tumour cells to sustain BCC growth and maintenance.

To determine whether vismodegib administration together with *Lgr5* lineage ablation can eliminate the LGR5-expressing drug-tolerant lesions that are responsible for tumour relapse, we administrated diphtheria toxin for five consecutive days in combination with vismodegib to *Krt14*^{CreER};*Ptch1*^{cKO};*Lgr5*$^{DTR-GFP}$ mice bearing persistent lesions (Extended Data Fig. 9h). *Lgr5* ablation combined with vismodegib administration led to almost total (99.5%) elimination of the persistent LGR5-expressing tumour cells (Extended Data Fig. 9i–k). We did not observe reappearance of LGR5$^+$ cells from the vast majority (94%) of the initial LGR5$^+$ persistent tumorigenic lesions 15 days after discontinuation of treatment with diphtheria toxin and vismodegib (Extended Data Fig. 9i, k, l), whereas HFSCs were replenished by LGR5-expressing cells as previously reported[22], indicating that there is little plasticity within the LGR5$^-$LRIG1$^+$ BCC cells to revert to LGR5$^+$ tumour cells after treatment with diphtheria toxin and vismodegib. The therapeutic benefit of *Lgr5* ablation in BCC is reminiscent of the effect of *Lgr5* ablation in a mouse model of colorectal cancer, in which *Lgr5* ablation prevents metastasis, and in human colorectal cancer organoids, in which *Lgr5* ablation promotes tumour regression and synergises with chemotherapy[23,24].

Lgr5 has been identified as a Wnt target gene, and acts as a co-receptor for R-spondin, positively regulating the Wnt signalling pathway[11].

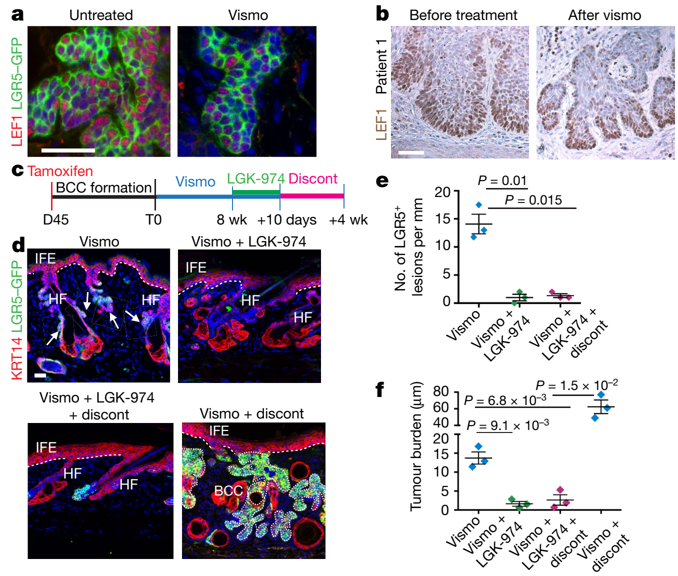


Fig. 4 | Dual Hh and Wnt inhibition eliminates vismodegib-persistent LGR5$^+$ tumour cells. a, Immunostaining for LGR5–GFP and LEF1 in untreated and vismodegib-treated *Ptch1*^{cKO} mice. **b**, Immunohistochemistry for LEF1 in biopsies from a patient before and after vismodegib treatment. **c**, Protocol for dual Hh and Wnt inhibition followed by treatment discontinuation. **d**, Immunostaining for LGR5–GFP and KRT14 upon vismodegib administration, dual inhibition of Wnt and Hh pathways and following discontinuation in *Ptch1*^{cKO}-derived BCCs. **e**, Number of LGR5$^+$ tumorigenic lesions per length of epidermis upon treatment and treatment discontinuation in *Ptch1*^{cKO}-induced BCCs (mean ± s.e.m.; $n = 3$ mice, 3 mm of skin analysed per mouse). Two-sided t-test. **f**, Quantification of the tumour burden upon treatment and treatment discontinuation in mice with *Ptch1*^{cKO}-induced BCCs (mean ± s.e.m.; $n = 3$ mice). See Source Data. Two-sided t-test. Three independent experiments per condition were analysed showing similar results (**a**) and two technical replicates were performed for each sample showing similar results (**b**). Hoechst nuclear staining in blue; scale bars, 50 µm. Dashed line delineates basal lamina; arrows indicate vismodegib-persistent lesions.

Administration of vismodegib decreased but did not abolish the expression of different members of the Wnt signalling pathway (Fig. 3c). Immunostaining for LEF1, a transcription factor that relays Wnt signalling and is a Wnt target gene in BCCs[9], and ISH for *Axin2*, another Wnt target gene, showed that both LEF1 and *Axin2* were expressed in LGR5$^+$ persistent lesions from mice and humans (Fig. 4a, b, Extended Data Fig. 10a, b), indicating that LGR5$^+$ persistent tumour cells are characterized by active Wnt signalling.

To assess whether dual Wnt and Hh inhibition can promote the elimination of LGR5$^+$ persistent tumour cells, we administered LGK-974, a porcupine Wnt inhibitor[25], and vismodegib for 10 consecutive days to *Ptch1*^{cKO} mice bearing LGR5$^+$ persistent lesions (Fig. 4c). Combined Wnt and Hh inhibition resulted in the disappearance of LEF1 expression consistent with efficient Wnt inhibition, the elimination of the vast majority (93%) of initial LGR5$^+$ drug-tolerant lesions and a substantial (87%) decrease in the tumour burden compared to vismodegib treatment alone (Fig. 4d–f, Extended Data Fig. 10c). We found no significant reduction in tumour burden after administration of the Wnt inhibitor alone, showing that although Wnt inhibition can block BCC initiation[9,14] it is not efficient as a monotherapy to induce clinically relevant BCC regression (Extended Data Fig. 10d–f), We then investigated whether rare residual tumour cells could lead to tumour relapse upon discontinuation of dual Wnt and Hh inhibition. Four weeks after discontinuation, which corresponds to the time that it takes for drug-tolerant lesions to regrow to their initial size upon vismodegib discontinuation, no tumour relapse was observed, as shown by the stable number of LGR5$^+$ tumour lesions and tumour burden (Fig. 4d–f).

Together, these results show that the synergy between Hh and Wnt inhibition in BCC leads to the elimination of the vast majority of LGR5$^+$ persistent tumour cells and thereby prevents tumour relapse.

In summary, we have shown that vismodegib induces BCC regression by promoting tumour differentiation and have identified a quiescent tumour cell population expressing LGR5 that persists after vismodegib treatment in different mouse models and human patients, promoting BCC relapse upon treatment discontinuation (Extended Data Fig. 11). The non-genetic mechanism of drug resistance described here differs from the previously described mutations in *Smo* or other genes that render cells insensitive to vismodegib treatment[6,7,19,20]. Administration of vismodegib promotes a switch from a proliferative state that fosters tumour growth to a tumour state characterized by Hh inhibition and slow-cycling properties that is fully reversible upon drug withdrawal and re-inducible upon a new cycle of vismodegib treatment. These persistent LGR5$^+$ tumour cells present residual Wnt signalling activity in both mouse and human BCCs and could be eliminated by dual Wnt and Hh inhibition, leading to tumour eradication in the majority of BCCs (Extended Data Fig. 11). Dual Wnt and Hh inhibition constitutes a clinically relevant strategy to avoid BCC relapse that might also be effective against other cancers, such as medulloblastoma, that are characterized by activation of Hh and Wnt signalling[26].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0603-3.

Received: 3 July 2017; Accepted: 16 August 2018;
Published online 8 October 2018.

1. Epstein, E. H. Basal cell carcinomas: attack of the hedgehog. *Nat. Rev. Cancer* **8**, 743–754 (2008).
2. Basset-Seguin, N., Sharpe, H. J. & de Sauvage, F. J. Efficacy of Hedgehog pathway inhibitors in basal cell carcinoma. *Mol. Cancer Ther*. **14**, 633–641 (2015).
3. Sekulic, A. et al. Efficacy and safety of vismodegib in advanced basal-cell carcinoma. *N. Engl. J. Med*. **366**, 2171–2179 (2012).
4. Kasper, M., Jaks, V., Hohl, D. & Toftgård, R. Basal cell carcinoma - molecular biology and potential new therapies. *J. Clin. Invest*. **122**, 455–463 (2012).
5. Tang, J. Y. et al. Inhibition of the hedgehog pathway in patients with basal-cell nevus syndrome: final results from the multicentre, randomised, double-blind, placebo-controlled, phase 2 trial. *Lancet Oncol*. **17**, 1720–1731 (2016).
6. Atwood, S. X. et al. Smoothened variants explain the majority of drug resistance in basal cell carcinoma. *Cancer Cell* **27**, 342–353 (2015).
7. Sharpe, H. J. et al. Genomic analysis of smoothened inhibitor resistance in basal cell carcinoma. *Cancer Cell* **27**, 327–341 (2015).
8. Youssef, K. K. et al. Identification of the cell lineage at the origin of basal cell carcinoma. *Nat. Cell Biol*. **12**, 299–305 (2010).
9. Youssef, K. K. et al. Adult interfollicular tumour-initiating cells are reprogrammed into an embryonic hair follicle progenitor-like fate during basal cell carcinoma initiation. *Nat. Cell Biol*. **14**, 1282–1294 (2012).
10. Brown, J. A. et al. TGF-β-induced quiescence mediates chemoresistance of tumor-propagating cells in squamous cell carcinoma. *Cell Stem Cell* **21**, 650–664.e8 (2017).
11. Barker, N., Tan, S. & Clevers, H. Lgr proteins in epithelial stem cell biology. *Development* **140**, 2484–2494 (2013).
12. Tian, H. et al. A reserve stem cell population in small intestine renders Lgr5-positive cells dispensable. *Nature* **478**, 255–259 (2011).
13. Page, M. E., Lombard, P., Ng, F., Göttgens, B. & Jensen, K. B. The epidermis comprises autonomous compartments maintained by distinct stem cell populations. *Cell Stem Cell* **13**, 471–482 (2013).
14. Yang, S. H. et al. Pathological responses to oncogenic Hedgehog signaling in skin are dependent on canonical Wnt/β3-catenin signaling. *Nat. Genet*. **40**, 1130–1135 (2008).
15. Rhee, H., Polak, L. & Fuchs, E. Lhx2 maintains stem cell character in hair follicles. *Science* **312**, 1946–1949 (2006).
16. Blanpain, C., Lowry, W. E., Geoghegan, A., Polak, L. & Fuchs, E. Self-renewal, multipotency, and the existence of two cell populations within an epithelial stem cell niche. *Cell* **118**, 635–648 (2004).
17. Latil, M. et al. Cell-type-specific chromatin states differentially prime squamous cell carcinoma tumor-initiating cells for epithelial to mesenchymal transition. *Cell Stem Cell* **20**, 191–204.e5 (2017).
18. Sánchez-Danés, A. et al. Defining the clonal dynamics leading to mouse skin tumour initiation. *Nature* **536**, 298–303 (2016).
19. Zhao, X. et al. A transposon screen identifies loss of primary cilia as a mechanism of resistance to SMO inhibitors. *Cancer Discov*. **7**, 1436–1449 (2017).
20. Whitson, R. J. et al. Noncanonical hedgehog pathway activation through SRF-MKL1 promotes drug resistance in basal cell carcinomas. *Nat. Med*. **24**, 271–281 (2018).
21. Eberl, M. et al. Tumor architecture and notch signaling modulate drug response in basal cell carcinoma. *Cancer Cell* **33**, 229–243.e4 (2018).
22. Hoeck, J. D. et al. Stem cell plasticity enables hair regeneration following Lgr5$^+$ cell loss. *Nat. Cell Biol*. **19**, 666–676 (2017).
23. de Sousa e Melo, F. et al. A distinct role for Lgr5$^+$ stem cells in primary and metastatic colon cancer. *Nature* **543**, 676–680 (2017).
24. Shimokawa, M. et al. Visualization and targeting of LGR5$^+$ human colon cancer stem cells. *Nature* **545**, 187–192 (2017).
25. Liu, J. et al. Targeting Wnt-driven cancer through the inhibition of Porcupine by LGK974. *Proc. Natl Acad. Sci. USA* **110**, 20224–20229 (2013).
26. Northcott, P. A. et al. Medulloblastomics: the end of the beginning. *Nat. Rev. Cancer* **12**, 818–834 (2012).

Acknowledgements We thank J.-M. Vanderwinden and M. Martens for help with confocal microscopy. C.B. is an investigator of WELBIO. A.S.-D. and J.-C.L. are supported by fellowships from the FNRS and FRIA, respectively. This work was supported by the FNRS, the Marian Family, the ULB fondation, the foundation Baillet Latour, and a consolidator grant from the European Research Council.

Author contributions A.S.-D. and C.B. designed the experiments, performed data analysis and wrote the manuscript; A.S.-D. performed most of the biological experiments; J.-C.L, G.L. and V.S. helped with *Lgr5* ablation experiments; M.L. performed immunostaining; C.D. performed FACS; E.M.-C., M.S., V.d.M. and J.T. provided patient samples; and A.B. performed GSEA.

Competing interests C.B. is a consultant at Genentech (San Francisco, USA).

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0603-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0603-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Ethical compliance. This study complied with all relevant ethical regulations regarding experiments involving mouse and human skin samples. Mouse colonies were maintained in a certified animal facility in accordance with European guidelines. Experiments involving mice presented in this work were approved by Comité d'Ethique du Bien Être Animal (Université Libre de Bruxelles) under protocol numbers 483N and 632N, which state that animals should be euthanized if they present tumours that exceed 1 cm in diameter. The BCCs observed in this study were microscopic and ranged from 1.5 mm to 100 μm in diameter; in none of the experiments performed did the BCCs exceed the limit (1 cm in diameter) described in protocols 483N and 632N.

Experiments involving human samples presented in this study were approved by the ethics committee of Vall d'Hebron Institute of Oncology (VHIO) and by the ethics committee of Hôpital Erasme under protocol number P2012/332. Permission and informed consent were obtained from all the patients in order to use their biopsies in this study.

Mice. *Krt14^{CreER}* transgenic mice[27] were kindly provided by E. Fuchs, Rockefeller University, USA. *Ptch1$^{fl/fl}$* mice[28] and *Rosa$^{SmoM2-YFP}$* mice[29] were obtained from the JAX repository. *Lgr5$^{DTR-GFP}$* mice (knockin mice that contain the diphtheria toxin receptor (DTR) fused to an enhanced green fluorescent protein (GFP) under the control of the *Lgr5* regulatory region, allowing us to identify LGR5-expressing cells using the GFP reporter and to selectively ablate *Lgr5* tumour cells by diphtheria toxin (DT) administration[12]) were kindly provided by Genentech (San Francisco, USA). *Tp53$^{fl/fl}$* mice[30] were obtained from the National Cancer Institute at Frederick.

Female and male animals were used for all experiments and equal gender ratios were respected in the majority of the analysis. Analysis of the different mutant mice was not blind and sample size was calculated to reach statistical significance. The experiments were not randomized.

Tumour induction. For *Ptch1^{cKO}* deletion, *Krt14^{CreER};Ptch1$^{fl/fl}$;Lrg5$^{DTR-GFP}$* mice and *Krt14^{CreER};Ptch1$^{fl/fl}$;Trp53$^{fl/fl}$;Lrg5$^{DTR-GFP}$* mice (1.5 months old) received one intraperitoneal injection of 2.5 mg tamoxifen on each of three consecutive days. For SmoM2 expression, *Krt14^{CreER};Rosa^{SmoM2};Lgr5$^{DTR-GFP}$* mice (1.5 months old) received one intraperitoneal injection of 1 mg tamoxifen. In the clonal induction experiments, *Krt14^{CreER};Rosa^{SmoM2}* mice (1.5 months old) received one intraperitoneal injection of 0.1 mg tamoxifen.

Vismodegib and LGK-974 administration. Vismodegib (GDC-0449) was kindly provided by Genentech (San Francisco, US) and LGK-974 was kindly provided by Novartis (Bâle, Switzerland). During vismodegib treatment, mice received 150 mg/kg vismodegib by oral gavage daily. Vismodegib was administered in two doses (one every 12 h).

During the 10-day LGK-974 treatment mice received: on each of the first six days, 10 mg/kg LGK-974 by oral gavage; and on each of the last four days one topical application of 100 μl of 2 mg/ml LGK-974 diluted in propylene glycol: ethanol (7:3 v/v). For oral gavage, LGK-974 and vismodegib were dissolved in 0.5% methylcellulose solution containing 0.2% Tween-80.

TPA and retinoic acid administration. TPA and retinoic acid (RA) were used to promote epidermal proliferation[31,32]. TPA (200 μl of 0.02 mg/ml solution in dimethyl sulfoxide) or retinoic acid (200 μl of 0.5 mM all-trans-RA (Sigma) in dimethyl sulfoxide) was administered daily to shaved mouse back skin for 2 weeks.

Diptheria toxin administration. For *Lgr5* lineage cell ablation, mice received a daily intraperitoneal injection of 50 μg/kg diphtheria toxin (Sigma).

Immunostaining in sections. The tail for the *SmoM2* model and ventral skin or back skin for *Ptch1^{cKO}* model were embedded in optimal cutting temperature compound (OCT, Sakura) and cut into 5–8 μm frozen sections using a CM3050S Leica cryostat (Leica Microsystems).

Immunostaining was performed on frozen sections. Owing to the fusion of SMOM2 with YFP and DTR with GFP, SMOM2-expressing and LGR5-expressing cells were detected using anti-GFP antibodies. Frozen sections were dried and then fixed with 4% paraformaldehyde/PBS (PFA) for 10 min at room temperature and blocked with blocking buffer for 1 h (PBS, horse serum 5%, BSA 1%, Triton 0.1%). Skin sections were incubated with primary antibodies diluted in blocking buffer overnight at 4 °C, washed with PBS for 3 × 5 min, and then incubated with Hoechst solution and secondary antibodies diluted in blocking buffer for 1h at room temperature. Finally, sections were washed with PBS for 3 × 5 min at room temperature and mounted in DAKO mounting medium supplemented with 2.5% Dabco (Sigma). Primary antibodies used were the following: anti-β4-integrin (rat, 1:200, BD, clone346-11A, ref. 553745, lot 5239648), anti-GFP (chicken, 1:3,000, Abcam, ref. ab13970, lot 236651-23), anti-active Caspase-3 (rabbit, 1:600, R&D, ref. AF835, lot CF23517031), anti-Ki67 (rabbit, 1:1,000, Abcam, ref. ab15580, lot GR3198193-1), anti-LRIG1 (goat, 1:500, R&D, ref. AF3688, lot ZPH0217111), anti-LEF1 (rabbit, 1:100, Cell Signaling, ref. 2230), anti-LHX2 (goat, 1:500, Santa Cruz, sc-19344, lot K1615), anti-CUX1 (rabbit,1:6,000, Santa Cruz, sc-13024), anti-TBX1 (rabbit, 1:100, Invitrogen), anti-ALCAM (goat, 1:1,000, Novus, ref. FAB1172F, lot AASW0111121), anti-KRT10 (rabbit, 1:3,000, Covance, ref. PRB-159P-0100), anti-KRT1 (rabbit, 1:3,000, Covance, ref. PRB-165P-0100), anti-KRT14 (rabbit, 1:3,000, Thermofisher), anti-SCD1 (goat, 1:500, Santa Cruz, ref. sc14719,lot H2610), anti-adipophilin (guinea pig, 1:5,000, Fitzgerald, ref. 20R-AP002, lot P17030911), anti-BrdU (mouse, 1:200, BD, clone 3D4, ref. 560209, lot 4293550), anti-MKL1 (rabbit, 1/200, Sigma, ref. HPA030782, lot C106712) and anti-ARL13B (rabbit, 1:2,000, ref. 17711-1-AP, Proteintech, lot 49885). The following secondary antibodies were used: anti-rabbit, anti-rat, anti-goat, anti-guinea pig and anti-chicken, conjugated to AlexaFluor488 (Molecular Probes) and to rhodamine Red-X and Cy5 (JacksonImmunoResearch). Images of the immunostained sections were acquired using an Axio Imager M2 microscope and Axiovision 4.8.2 software (Carl Zeiss).

Immunostaining in whole-mounts. Whole-mounts of tail epidermis were performed as previously described[33] and used to quantify the proportion of surviving clones. Pieces of tail were incubated for 1 h at 37 °C in EDTA 20 mM in PBS on a rocking plate, then the dermis and epidermis were separated using forceps and the epidermis was fixed for 30 min in paraformaldehyde (PFA) 4% with agitation at room temperature and washed three times with PBS.

For the immunostaining, tail skin pieces were blocked with blocking buffer for 3 h (PBS, horse serum 5%, Triton 0.8%) on a rocking plate at room temperature. Next, the skin pieces were incubated with primary antibodies diluted in blocking buffer overnight at 4 °C. The next day, they were washed with PBS-Tween 0.2% for 3 × 10 min at room temperature, and then incubated with the secondary antibodies diluted in blocking buffer for 3h at room temperature, washed 2 × 10 min with PBS-Tween 0.2% and washed for 10 min in PBS. Finally, they were incubated in Hoechst diluted in PBS for 30 min at room temperature in the rocking plate, washed 3 × 10 min in PBS and mounted in DAKO mounting medium supplemented with 2.5% Dabco (Sigma). Primary antibodies used were the following: anti-GFP (rabbit, 1:100, BD, ref. A11122), anti-β4-integrin (rat, 1:200, BD, ref. 553745) and anti-KRT31 (guinea pig, 1:200, Progen, ref. GP-hHa1). The following secondary antibodies were used: anti-rabbit, anti-rat and anti-guinea pig, conjugated to AlexaFluor488 (Molecular Probes), to rhodamine Red-X (JacksonImmunoResearch) and to Cy5 (1:400, Jackson ImmunoResearch).

BrdU and EdU label retention studies. For the BrdU studies, mice received three daily intraperitoneal injections (150 μl of 10 mg/ml, every 8 h) for three consecutive days. For EdU studies, mice received three daily intraperitoneal injections (150 μl of 1 mg/ml, every 8 h) for three consecutive days. EdU and BrdU stainings were performed as described[18].

In situ hybridization and RNA FISH. The tail in the *SmoM2* model and ventral skin in the *Ptch1^{cKO}* model were embedded in OCT and cut into 5–8 μm frozen sections using a CM3050S Leica cryostat (Leica Microsystems). Samples were fixed for 30 min in 4% PFA at 4 °C and the in situ protocol was performed according to the manufacturer's instructions (Advanced Cell Diagnostics). The following mouse probes were used: Mm-Lgr5 cat. No. 312171, Mm-Gli1 cat. No. 311001-C2, Mm-Axin2 cat no.400331-C3, Mm-Defensinβ6 cat no.430141-C3

Human samples were fixed in 4% formalin and embedded in paraffin. Cut sections were deparaffinized and rehydrated before proceeding to the in situ hybridization, which was performed according to the manufacturer's instructions. The following probes were used: Hs-Lgr5-C2 cat. no. 310991-C2, Hs-Lgr5 cat. no. 311021 and Hs-Axin2 cat no.400241-C3.

A confocal microscope (LSM-780, Carl Zeiss) and ZEN 2.3 software were used to acquire and analyse the ISH images.

Immunohistochemistry. For KRT14, Ki67, KRT10 and LEF1 immunohistochemistry in human samples, paraffin sections were deparaffinized and rehydrated, followed by antigen unmasking performed for 20 min at 98 °C in citrate buffer (pH 6) using the PT module. Endogenous peroxidase was blocked using 3% H_2O_2 (Merck) in methanol for 10 min at room temperature. Endogenous avidin and biotin were blocked using the Endogenous Blocking kit (Invitrogen) for 20 min at room temperature. Nonspecific antigen blocking was performed using blocking buffer. Mouse anti-KRT14 (rabbit, 1:2,000, Thermofisher), anti-Ki67 (rabbit, 1:400, Abcam, ab15580), anti-KRT10 (rabbit, 1:200, Biolegend, ref. 90541) and anti-LEF1 (rabbit, 1:100, Cell Signaling, ref. 2230) were incubated overnight at 4 °C. Anti-rabbit biotinylated with blocking buffer, standard ABC kit, and ImmPACT DAB (Vector Laboratories) was used for the detection of horseradish peroxidase (HRP) activity. Slides were then dehydrated and mounted using SafeMount (Labonord).

FACS isolation of tumour cells and microarray analysis. Isolation of tumour cells was performed as previously described[34]. In brief, *Lgr5$^{DTR-GFP}$* and *Krt14^{CreER};Ptch1$^{fl/fl}$;Lgr5$^{DTR-GFP}$* mice untreated and upon 8 weeks of vismodegib treatment were killed by decapitation. Back skin was placed in a Petri dish and a sterile scalpel was used to remove the adipose tissue and muscle. The skin tissue was incubated with thermolysin (Sigma) for 1 h at 37 °C and then a scalpel was used to separate epidermis from the dermis. The epidermal tissue was chopped into pieces and resuspended in PBS supplemented with 5% chelated fetal calf serum and filtered with 70 μm and 40 μm cell strainers (BD). Cells were stained using

anti-LRIG1 (goat polyclonal, R&D Systems, AF3688) followed by the secondary antibody donkey anti-goat-Alexa 647 (Invitrogen).

LRG5$^+$LRIG1$^+$ and LGR5$^-$LRIG1$^+$ cells from untreated or vismodegib-treated (8 weeks) *Krt14^{CreER};Ptch1$^{fl/fl}$;Lgr5$^{DTR-GFP}$* mice were sorted using LRIG1 staining and native LGR5–GFP. Two thousand sorted cells per sample were collected directly in 45 μl lysis buffer (20 mM DTT, 10 mM Tris–HCl pH 7.4, 0.5% SDS, 0.5 μg μl−1 proteinase K). Samples were then lysed at 65 °C for 15 min and frozen. RNA isolation, amplification and microarray were performed at the IRB Functional Genomics Core, Barcelona. cDNA synthesis, library preparation and amplification were performed as described[35]. Microarrays using Mouse Genome 430pm strip Affymetrix array were performed and the data were normalized using RMA algorithm. Biological duplicates were performed for all conditions. Genetic signatures were obtained by considering genes presenting a fold change greater or smaller than 2 or −2, respectively, in each replicate.

FACS isolation of CD71$^+$ and CD71$^-$ populations of tumour cells, RNA extraction and quantitative PCR. Isolation of tumour cells from mouse skin was performed as described above. Cells were stained using anti-LRIG1 (goat polyclonal, R&D Systems, AF3688) and anti-CD71–PE (rat, BD Biosciences, 553267) followed by the secondary antibody donkey anti-goat-Alexa 647 (Invitrogen). Seven thousand FACS-sorted cells were collected directly in the lysis buffer provided by the manufacturer (RNAeasy Microkit, Quiagen) and RNA extraction was then carried out according to the manufacturer's protocol. Purified RNA was used to synthesize the first-strand complementary DNA using SuperScript II (Invitrogen) with random hexamers (Roche). Quantitative PCR analyses were carried out with Light Cycler 96 (Roche). Primers used: Ki67-F: CCTGCCTCAGATGGCTCAAA, Ki67-R: GGTTCCCTGTAACTGCTCCC, Aurka-F: AACACAACGCAAGCCA AAGG, Aurka-R: GGCCAGTTGGAGGTTTGGAA, Krt10-F: AACTGACAAT GCCAACGTGC, Krt10-R: TAGGTAG GCCAGCTCTTCGT, Krt1-F: ACAACCCGGACCCAAAACTT, Krt1-R: CT CTGCGTTGGTCCTCTTGT, Scd1-F: ACACCATGGCGTTCCAGAAT and Scd1-R: AGCTTCTCGGCTTT CAGGTC. Normalizers: HPRT-F: GCAGTA CAGCCCCAAAATGG, HPRT-R: TCCAACAAAGTCTGGCCTGT, βActin-F: GAAGCTGTGCTATGTTGCTCTA, βActin-R: CAATAGTGATGACCTGGCCGT, β2M-F: TCACCCCCACTGAGAC TGAT, β2M-R: TCCCAGTAGACGGTCTTGGG, Gapdh-F: CGTGTTCCTACCC CCAATGT, Gapdh-R: GTGTAGCCCAAGATGCCC TT, Tbox-F: GTACCGCAGC TTCAAAAT ATTGTAT and Tbox-R: AAATCAACGCAGTTGTGCGTG

Sequencing of the *Smo* gene in vismodegib-persistent lesions. A total of 200,000 LGR5$^+$LRIG1$^+$ tumour cells from three *Krt14^{CreER};Ptch1^{cKO};Lgr5$^{DTR-GFP}$* mice treated for 8 weeks with vismodegib were FACS sorted following the protocol described above. Exons 3–12 of the mouse *Smo* gene were amplified using PCR and the products of the PCR were purified using the Monarch DNA Gel Extraction Kit (ref. T1020). The products of the PCR were sequenced following the Sanger standard using chemistry BigDy31.1, the cycle sequencing technology based on dideoxy chain termination/cycle sequencing and performed on a ABI 3730XL sequencer. SnapGene version 4.1.3 was used for the analysis. Information of the amplification primers and sequencing results can be found in Source Data.

Grafting experiments. For transplantation experiments, 100,000 cells that had been FACS-sorted to obtain pure populations of LGR5$^+$LRIG1$^+$ and LGR5$^-$LRIG1$^+$ cells were transplanted into the interscapular fat pad of NOD-SCID immunodeficient mice. The 100,000 LGR5$^+$LRIG1$^+$ and LGR5$^-$LRIG1$^+$ cells were mixed in a proportion of 1/40 with tumour-associated fibroblasts from the same tumours (FACS-sorted using CD140a marker (clone APA5; eBiosciences)). The tumour cells and fibroblasts were embedded in 50 μl matrigel containing ROCK inhibitor (3.3 μg/ml) and transplanted into the fat pad.

GSEA analysis. The GSEA program was downloaded from the BROAD institute website (http://www.broadinstitute.org/gsea/). We used the GSEA preranked option with standard parameters of weighted enrichment score calculation to run the GSEA against a user-supplied fold-change-ranked list of genes. The results of the enrichment analysis were plotted using R software

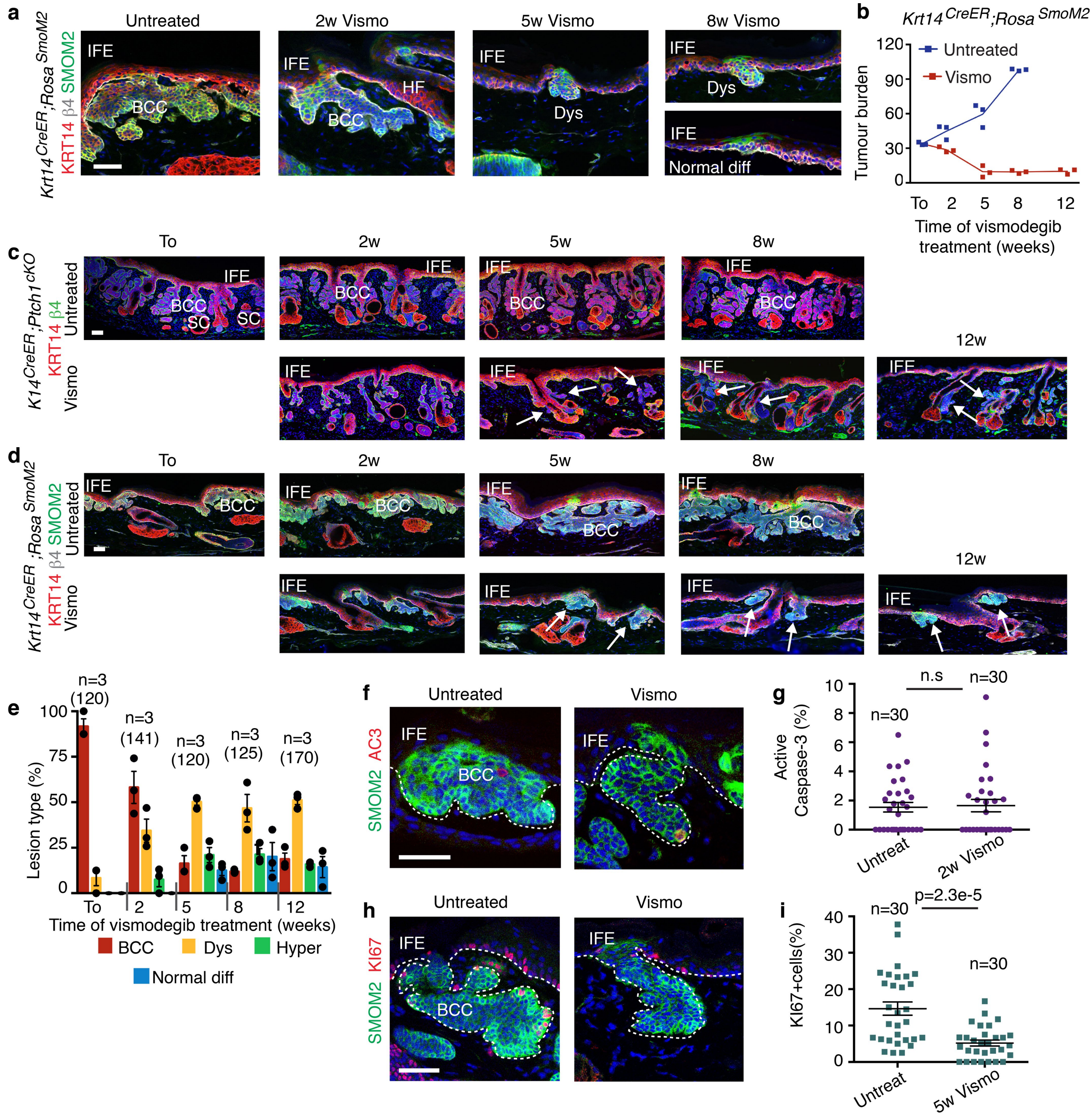
***Ptch1* deletion.** To determine the deletion of the two *Ptch1* alleles in the LGR5$^+$LRIG1$^+$ and LGR5$^-$LRIG1$^+$ populations, 200,000 cells were FACS-sorted and DNA was extracted using the QiAmp DNA Mini-Kit (Qiagen). The following primers were used to determine the presence of the floxed/floxed or deleted alleles: Forward: AAAGAGATCTTGTGGGCAAGG; Reverse: CTACTTCCATTTGTCACGTCC.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

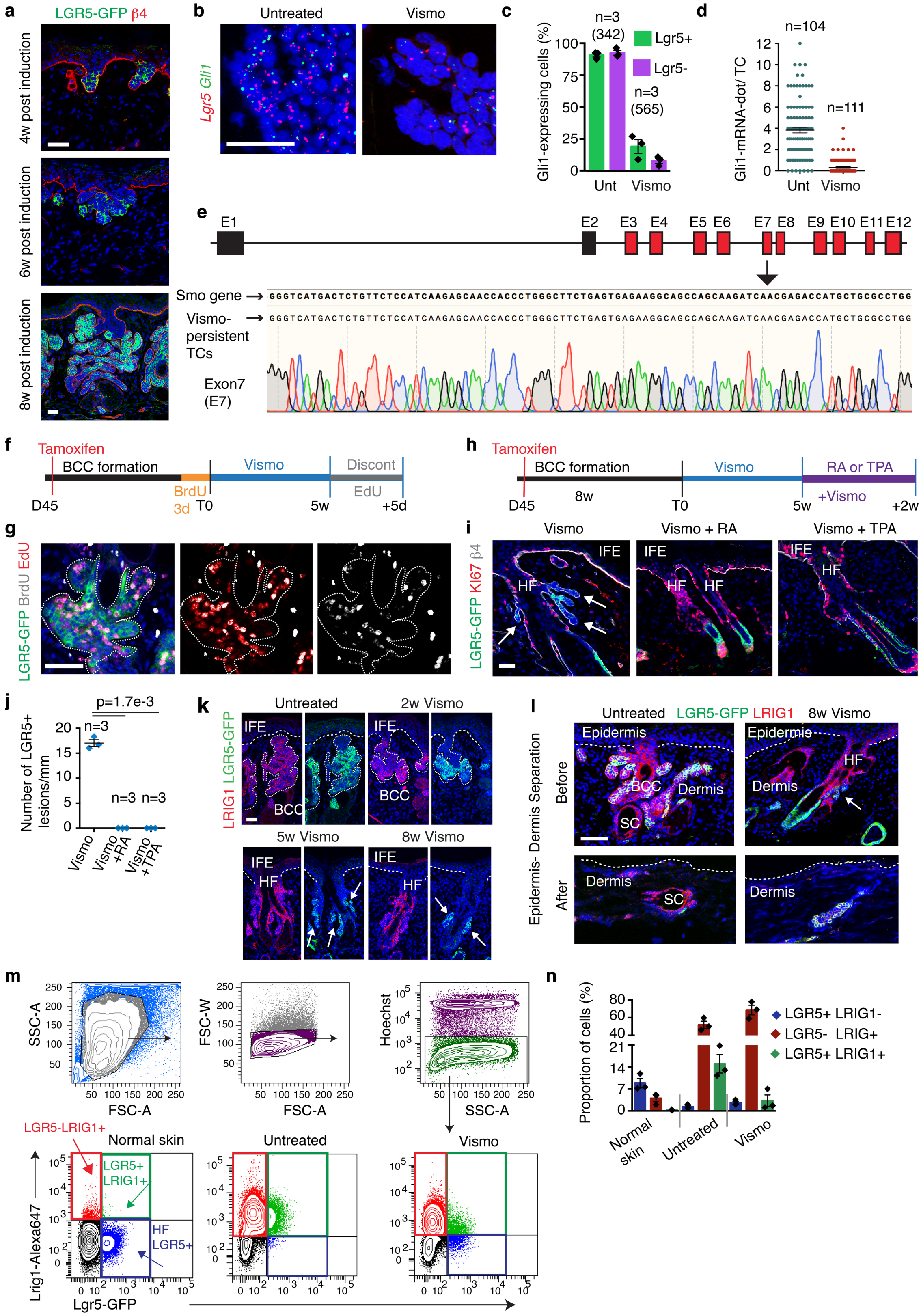
Data availability

Data associated with this study have been deposited in the NCBI Gene Expression Omnibus under accession number GSE117458 (microarray).

27. Vasioukhin, V., Degenstein, L., Wise, B. & Fuchs, E. The magical touch: genome targeting in epidermal stem cells induced by tamoxifen application to mouse skin. *Proc. Natl Acad. Sci. USA* **96**, 8551–8556 (1999).
28. Uhmann, A. et al. The Hedgehog receptor Patched controls lymphoid lineage commitment. *Blood* **110**, 1814–1823 (2007).
29. Mao, J. et al. A novel somatic mouse model to survey tumorigenic potential applied to the Hedgehog pathway. *Cancer Res.* **66**, 10171–10178 (2006).
30. Jonkers, J. et al. Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nat. Genet.* **29**, 418–425 (2001).
31. Aldaz, C. M., Conti, C. J., Gimenez, I. B., Slaga, T. J. & Klein-Szanto, A. J. Cutaneous changes during prolonged application of 12-*O*-tetradecanoylphorbol-13-acetate on mouse skin and residual effects after cessation of treatment. *Cancer Res.* **45**, 2753–2759 (1985).
32. Collins, C. A. & Watt, F. M. Dynamic regulation of retinoic acid-binding proteins in developing, adult and neoplastic skin reveals roles for β-catenin and Notch signalling. *Dev. Biol.* **324**, 55–67 (2008).
33. Braun, K. M. et al. Manipulation of stem cell proliferation and lineage commitment: visualisation of label-retaining cells in wholemounts of mouse epidermis. *Development* **130**, 5241–5255 (2003).
34. Jensen, K. B., Driskell, R. R. & Watt, F. M. Assaying proliferation and differentiation capacity of stem cells using disaggregated adult mouse epidermis. *Nat. Protoc.* **5**, 898–911 (2010).
35. Gonzalez-Roca, E. et al. Accurate expression profiling of very small cell populations. *PLoS One* **5**, e14418 (2010).

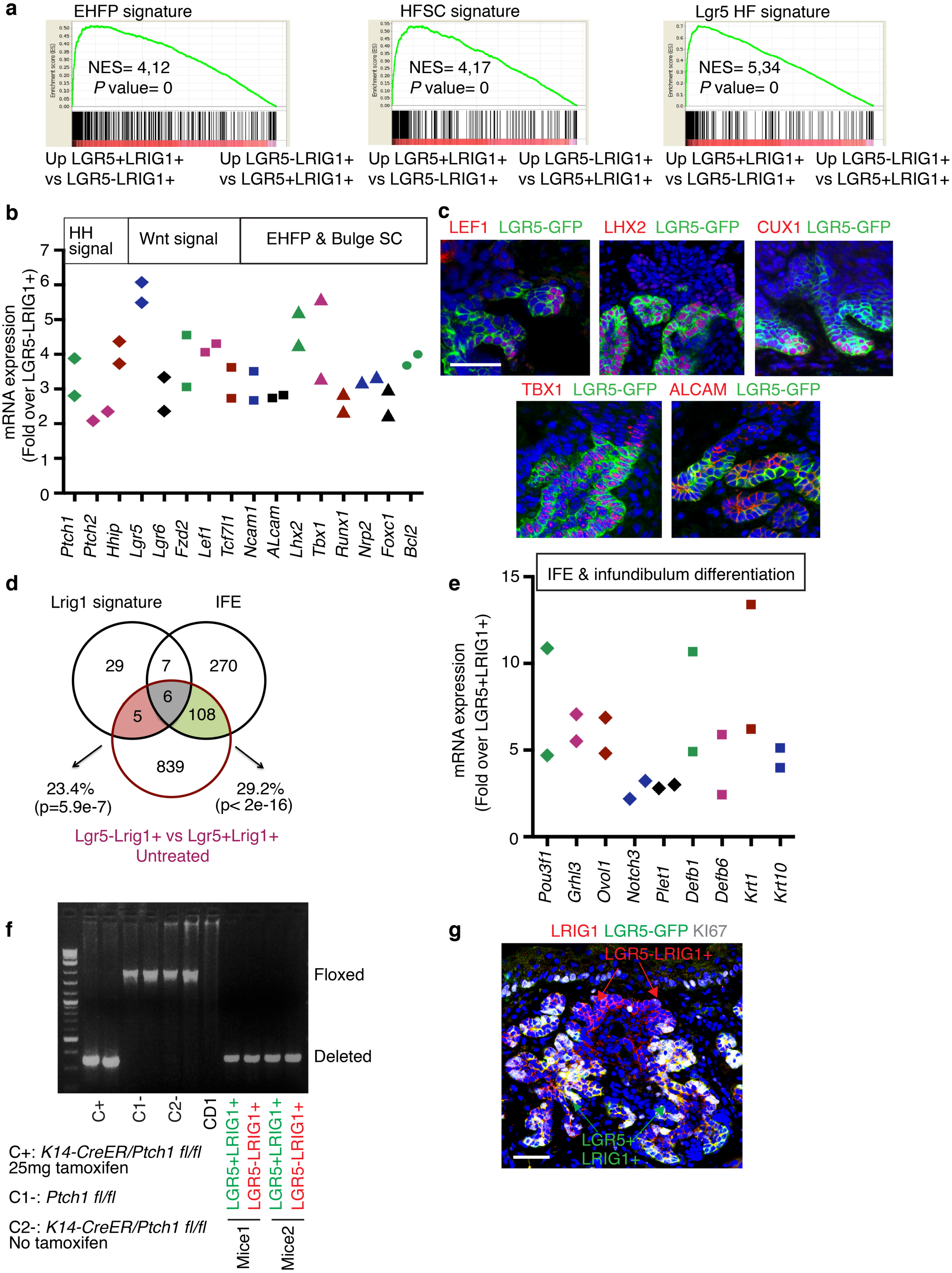


Extended Data Fig. 1 | Vismodegib leads to tumour shrinkage and emergence of vismodegib-persistent lesions in mice. a, Immunostaining for SMOM2, KRT14 and β4-integrin in tail from *SmoM2* mice after different durations of vismodegib administration. **b**, Tumour burden in micrometres (total area occupied by tumours divided by the length of the analysed epidermis) in untreated and vismodegib-treated *SmoM2* mice ($n = 3$ mice analysed per time point and condition). Centre values define the mean. See Source Data for description of the skin length and tumour area analysed per mouse. **c**, Immunostaining for KRT14 and β4-integrin in ventral skin from *Ptch1*^{cKO} mice. **d**, Immunostaining for SMOM2, KRT14 and β4-integrin in tail skin from *SmoM2* mice. **e**, Quantification (mean ± s.e.m.) of lesion type upon vismodegib treatment in *SmoM2* mice ($n = 3$ mice, total number of lesions analysed per time point indicated in parentheses). **f**, Immunostaining for active caspase-3 (AC3) and SMOM2. **g**, Percentage of AC3^{+} tumour cells (mean ± s.e.m.) in untreated and vismodegib-treated *SmoM2* mice ($n = 30$ lesions analysed from 3 mice). Two-sided *t*-test. **h**, Immunostaining for Ki67 and SMOM2. **i**, Percentage of Ki67^{+} tumour cells (mean ± s.e.m.) in untreated and vismodegib-treated *SmoM2* mice ($n = 30$ lesions analysed from 3 mice). Two-sided *t*-test. Three independent experiments per condition were analysed showing similar results (**a**, **c**, **d**, **f**, **h**). Hoechst nuclear staining in blue; scale bars, 100 μm (**c**, **d**), 50 μm (**a**, **f**, **h**). Dashed line delineates basal lamina. Arrows indicate vismodegib-persistent lesions.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Vismodegib-persistent lesions express LGR5 in mice. **a**, Immunostaining for LGR5–GFP and β4-integrin at different time points after tamoxifen administration in the *Ptch1*^{cKO} model. **b**, In situ hybridization for *Lgr5* and *Gli1* in untreated and treated tumour cells in *SmoM2* mice. **c**, Percentage (mean ± s.e.m.) of tumour cells (LGR5^{+} and LGR5^{-}) that express *Gli1* in *SmoM2* mice ($n = 3$ mice, total number of cells analysed indicated in parentheses). **d**, Distribution (mean ± s.e.m.) of the number of *Gli1* mRNA dots per tumour cell with and without treatment in *SmoM2* mice ($n = 104$ and 111 total tumour cells from 3 mice per condition and time point). **e**, Representation of the mouse *Smo* gene, showing in red the exons (E) in which genetic mutations have been described[6,7] (top). Results from the sequencing of exon 7 from vismodegib-persistent lesions obtained by pooling drug-persistent cells from three *Krt14*^{CreER};*Ptch1*^{cKO};*Lgr5*$^{DTR\text{-}GFP}$ mice, showing absence of genetic mutations in the exon analysed (bottom). See Source Data for the results of the sequencing of exons 3–12. **f**, Protocol for BrdU and EdU double-labelling studies in *Ptch1*^{cKO}-induced BCCs followed by vismodegib administration and discontinuation. **g**, Immunostaining for LGR5–GFP, BrdU and EdU in *Ptch1*cKO-derived BCCs following 5 days of vismodegib discontinuation. **h**, Protocol for treatment with vismodegib and retinoic acid (RA) or TPA. **i**, Immunostaining for LGR5–GFP, Ki67 and β4 in the back skin of *Ptch1*^{cKO} mice treated with vismodegib and RA or TPA. **j**, Quantification of LGR5^{+} tumorigenic lesions per length of skin upon treatment with vismodegib or vismodegib with RA or TPA ($n = 3$ mice, 3 mm of skin analysed per mouse). Two-sided *t*-test. **k**, Immunostaining for LGR5–GFP and LRIG1 in untreated and treated *Ptch1*^{cKO} mice. **l**, Immunostaining for LGR5–GFP and LRIG1 in untreated and vismodegib-treated (8 weeks) mice before and after enzymatic and physical separation of epidermis from dermis in *Krt14*^{CreER};*Ptch1*^{cKO};*Lgr5*$^{DTR\text{-}GFP}$ mice. Note that hair follicles co-expressing LGR5 and LRIG1 and sebaceous cysts remained in the dermal fraction whereas the BCCs were isolated with the epidermal fraction, indicating that normal hair follicles did not significantly contaminate the FACS-isolated tumour cells. **m**, Cell sorting strategy to isolate LGR5^{+}LRIG1^{-}, LGR5^{+}LRIG1^{+} and LGR5^{-}LRIG1^{+} in normal skin and in *Ptch1*^{cKO}-derived BCCs before and after vismodegib administration. Forward scatter (FSC) and side scatter (SSC) were performed to exclude cell debris and doublets. Living cells were selected by Hoechst dye exclusion. Finally, the different LGR5 and LRIG1 cell populations were isolated by FACS sorting. A, area; W, width. **n**, Proportion of cells (mean ± s.e.m.) expressing LGR5–GFP and LRIG1 determined by FACS ($n = 3$ independent experiments per condition). These experiments indicate that LRIG1 can be used to discriminate between LGR5^{+} cells from the HFSC or lower hair follicle (LGR5^{+}LRIG1^{-}) and BCC cells (LGR5^{+}LRIG1^{+}). Three independent experiments per condition were analysed with similar results (**a**, **k**, **l**). Hoechst nuclear staining in blue; scale bars, 50 μm (**a**, **i**, **k**, **l**) and 25 μm (**b**). Dashed line delineates basal lamina separating IFE from the dermis. Dotted line delineates BCC. Arrows indicate vismodegib-persistent lesions.



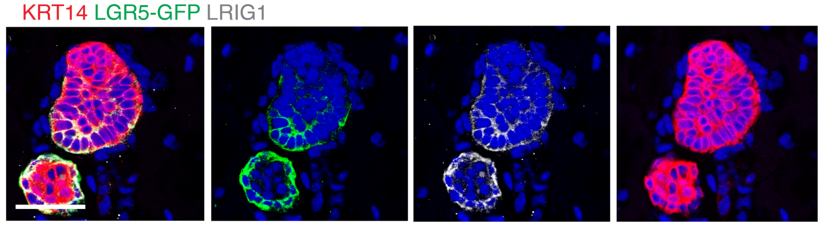
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Characterization of LGR5$^+$LRIG1$^+$ and LGR5$^-$LRIG1$^+$ tumour cells. a, GSEA showing the enrichment of genes upregulated in the LGR5$^+$LRIG1$^+$ population compared to the LRG5$^-$LRIG1$^+$ population from two independent microarray experiments with the EHFP[15] (left) in telogen HFSCs[16] (middle) and hair follicle Lgr5-expressing cell signatures[17] (right), showing that LGR5-expressing BCC cells express many genes of the embryonic and adult hair follicle signatures. The normalized enrichment score (NES) and *P* value (one-sided test) were calculated using the GSEA program. **b**, mRNA expression of genes upregulated in LGR5$^+$LRIG1$^+$ tumour cells compared to LGR5$^-$LRIG1$^+$ tumour cells in untreated conditions ($n = 2$ independent microarray experiments). **c**, Immunostaining for LGR5–GFP with LEF1, LHX2, CUX1, TBX1 and ALCAM in untreated *Ptch1^{cKO}*-derived BCCs. **d**, Venn diagram showing the similarities and differences between genes that were upregulated more than twofold from two independent microarray experiments in LGR5$^-$LRIG1$^+$ versus LGR5$^+$LRIG1$^+$ cells compared to IFE[16] and LRIG1[13] signatures. *P* value calculated using the hypergeometric test for each intersection of two subsets of genes with phyper function in R software. The high overlap indicates that LGR5$^-$LRIG1$^+$ cells expressed IFE and infundibulum differentiation markers. **e**, mRNA expression of genes upregulated in LGR5$^-$LRIG1$^+$ tumour cells compared to LGR5$^+$LRIG1$^+$ cells in untreated conditions ($n = 2$ independent microarray experiments). **f**, PCR analysis of the recombination of the floxed *Ptch1* alleles in control samples and in FACS-isolated tumour-derived LGR5$^+$LRIG1$^+$ and LGR5$^-$LRIG1$^+$ populations from *Ptch1^{cKO}*-induced BCCs. Two technical replicates were analysed for each sample with similar results. **g**, Immunostaining for LGR5–GFP, LRIG1 and Ki67 in *Ptch1^{cKO}*-derived BCCs shows higher proliferation rate in LGR5$^+$LRIG1$^+$ than in LGR5$^-$LRIG1$^+$ tumour cells. Three independent experiments per condition were analysed with similar results (**c**, **g**). Hoechst nuclear staining in blue; scale bars, 50 μm.

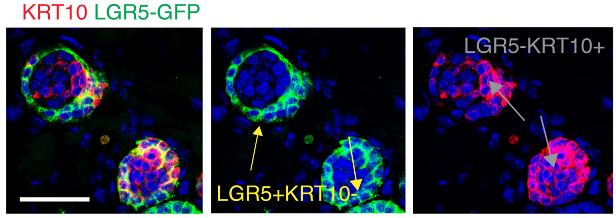
a

	Number grafted mice presenting KRT14+ cells
Krt14^{CreER};Ptch1^{cKO};Lgr5$^{DTR\text{-}GFP}$	
LGR5+LRIG1+	0/7
LGR5+LRIG1-	0/7
Krt14^{CreER};Ptch1^{cKO};Tp53^{cKO};Lgr5$^{DTR\text{-}GFP}$	
LGR5+LRIG1+	3/7
LGR5+LRIG1-	0/7

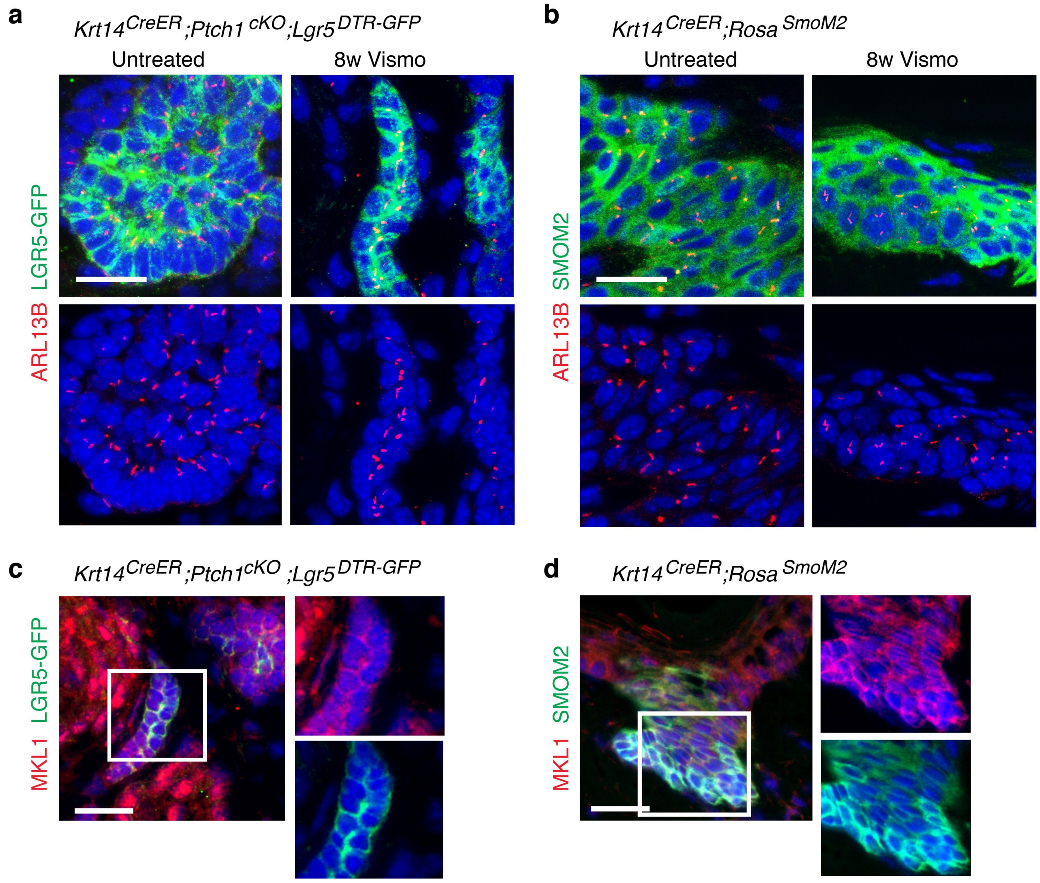
b KRT14 LGR5-GFP LRIG1



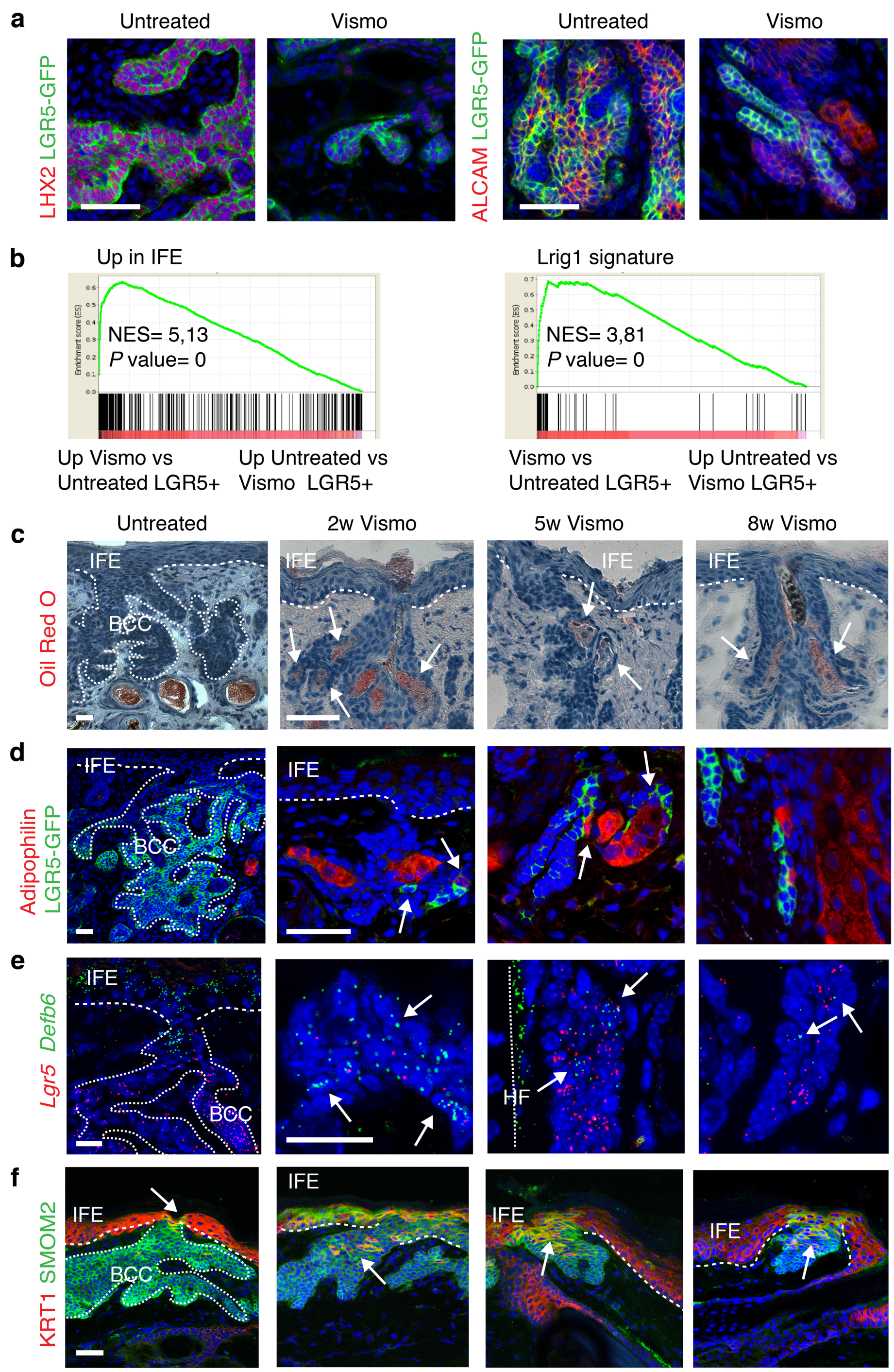
c KRT10 LGR5-GFP



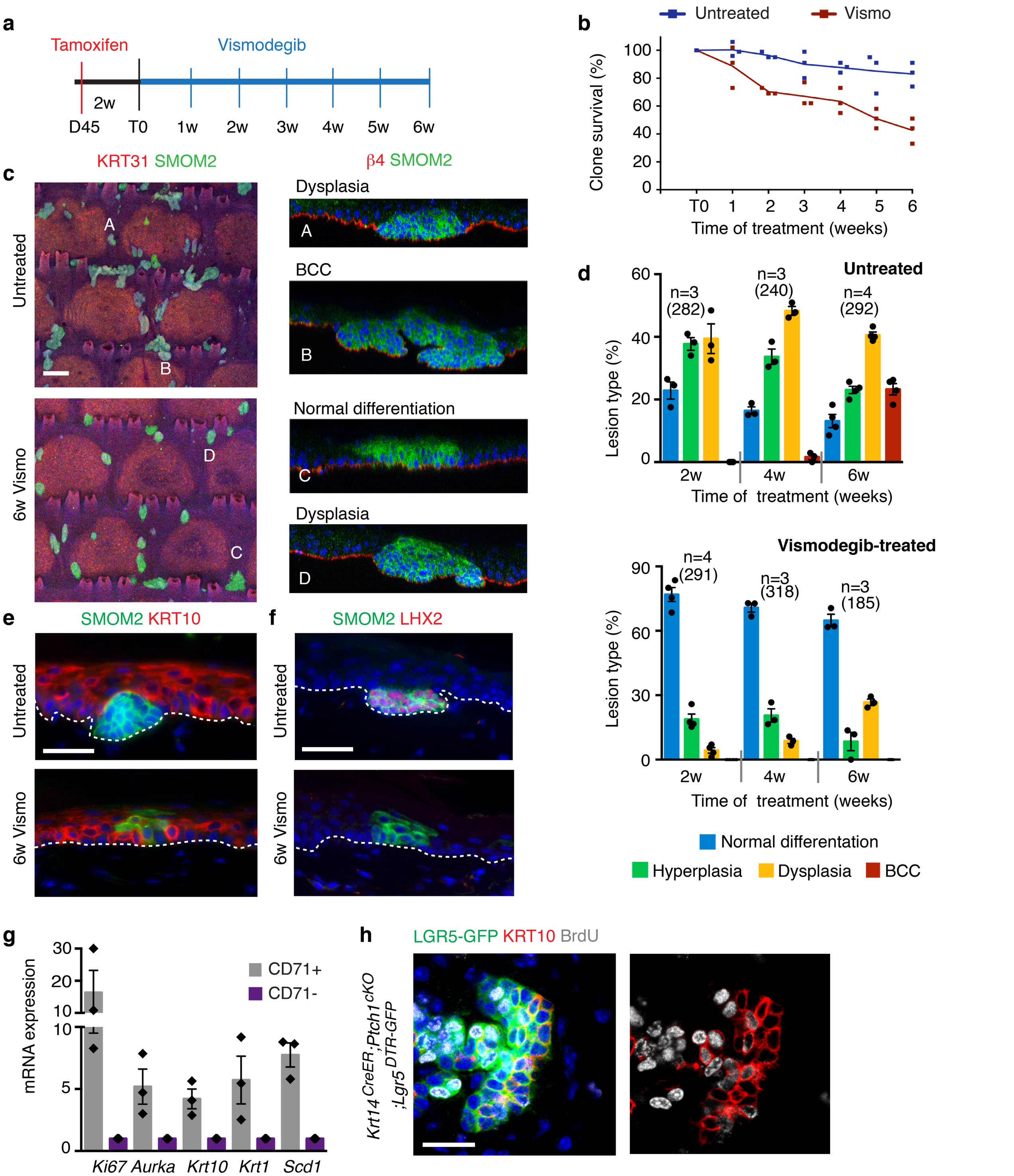
Extended Data Fig. 4 | Transplantation of LGR5$^+$LRIG1$^+$ *Ptch1;Trp53* double conditional knockout BCC cells leads to the formation of BCC-like structures. a, Table summarizing the number of grafted mice that presented KRT14$^+$ BCC-like structures upon transplantation of FACS-isolated LGR5$^+$LRIG1$^+$ and LGR5$^-$LRIG1$^+$ cells from BCCs arising in *Krt14^{CreER};Ptch1^{cKO};Lgr5$^{DTR–GFP}$* and *Krt14^{CreER};Ptch1^{cKO};Trp53^{cKO};Lgr5$^{DTR–GFP}$* mice. **b**, **c**, Immunostaining for LGR5–GFP, KRT14 and LRIG1 (**b**) and for LGR5–GFP and KRT10 (**c**) in the BCC-like structures obtained upon transplantation of LGR5$^+$LRIG1$^+$ cells from *Ptch1;Trp53* double conditional knockout BCCs in the dorsal fat pads of NOD/SCID mice. Three independent experiments per condition were analysed with similar results (**b**, **c**). Hoechst nuclear staining in blue; scale bars, 50 μm.



Extended Data Fig. 5 | Vismodegib-persistent lesions do not show decreased primary cilia numbers or nuclear localization of MKL1. **a**, **b**, Immunostaining for ARL13B and LGR5–GFP in *Ptch1*^{cKO} model (**a**) and for ARL13B and SMOM2 in *SmoM2* model (**b**) in untreated and vismodegib-treated lesions. **c**, **d**, Immunostaining for MKL1 and LGR5–GFP in vismodegib-persistent lesions in *Ptch1*^{cKO} mice (**c**) and for MKL1 and SMOM2 in vismodegib-persistent lesions in *SmoM2* mice (**d**) treated for 8 weeks with vismodegib. White boxes are expanded on right. Three independent experiments per condition were analysed with similar results. Scale bars, 25 μm.



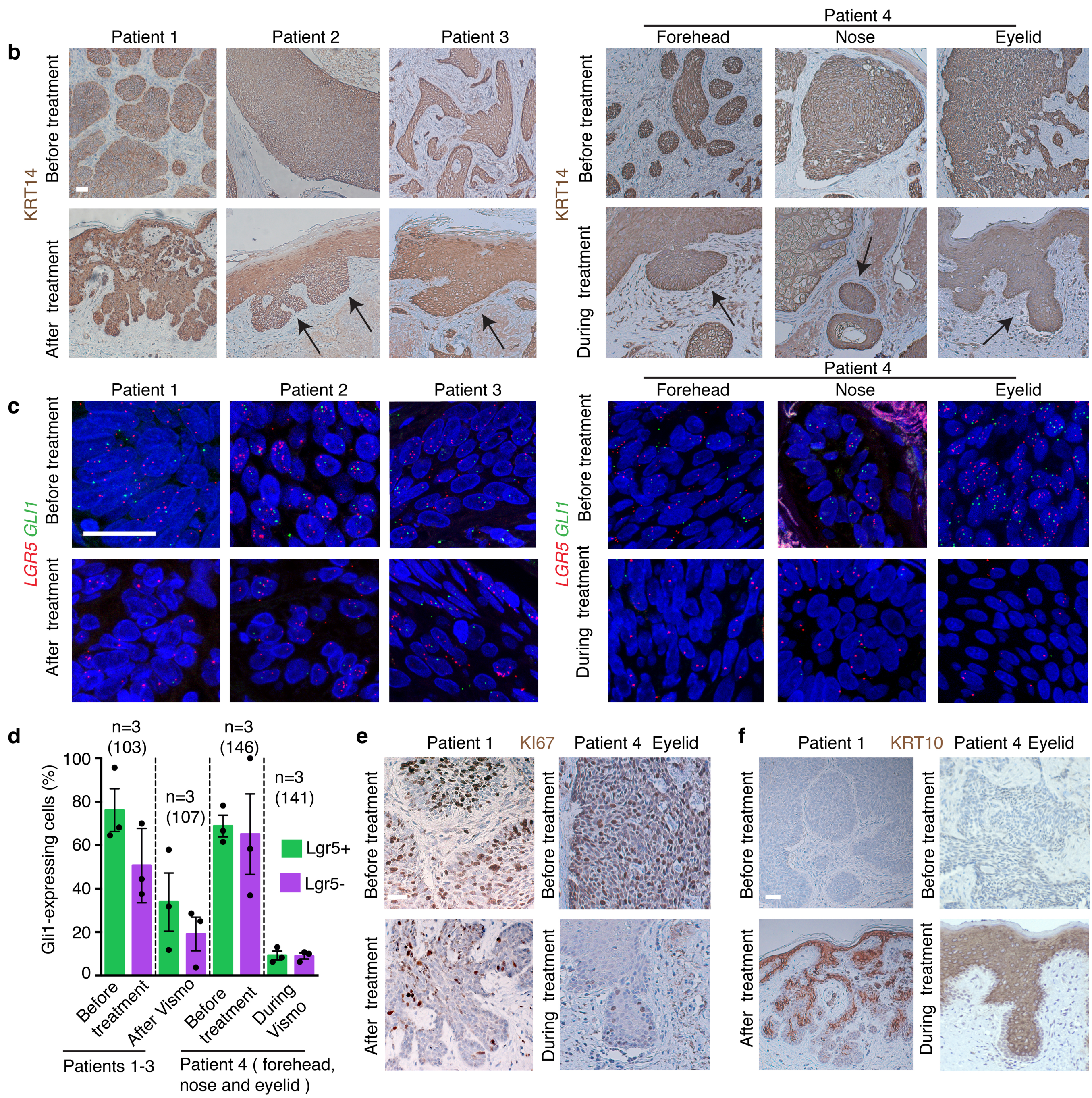
Extended Data Fig. 6 | Vismodegib promotes BCC differentiation. **a**, Immunostaining for LGR5–GFP, LHX2 and ALCAM in untreated and vismodegib-treated *Ptch1*^{cKO}-derived BCCs. **b**, GSEA showing enrichment of genes upregulated in LGR5^{+}LRIG1^{+} vismodegib-treated tumours compared to untreated BCCs with IFE[16] and LRIG1[13] signatures, showing that vismodegib treatment promotes the expression of the IFE and infundibulum signatures. The normalized enrichment score (NES) and *P* value (one-sided test) were calculated using the GSEA program. **c**, Oil Red O and haematoxylin and eosin staining in ventral skin of untreated and vismodegib-treated *Ptch1*^{cKO} mice. Arrows indicate areas of sebaceous differentiation. **d**, Immunostaining for LGR5–GFP and adipophilin in untreated and vismodegib-treated *Ptch1*^{cKO}-derived BCCs. Arrows indicate areas of sebaceous differentiation. **e**, In situ hybridization for *Lgr5* and *Defb6* in untreated and vismodegib-treated *Ptch1*^{cKO}-derived BCCs. **f**, Immunostaining for KRT1 and SMOM2 in untreated and vismodegib-treated *SmoM2* mice. Three independent experiments per condition were analysed with similar results (**a**, **c**–**f**). Hoechst nuclear staining in blue; scale bars, 50 µm.



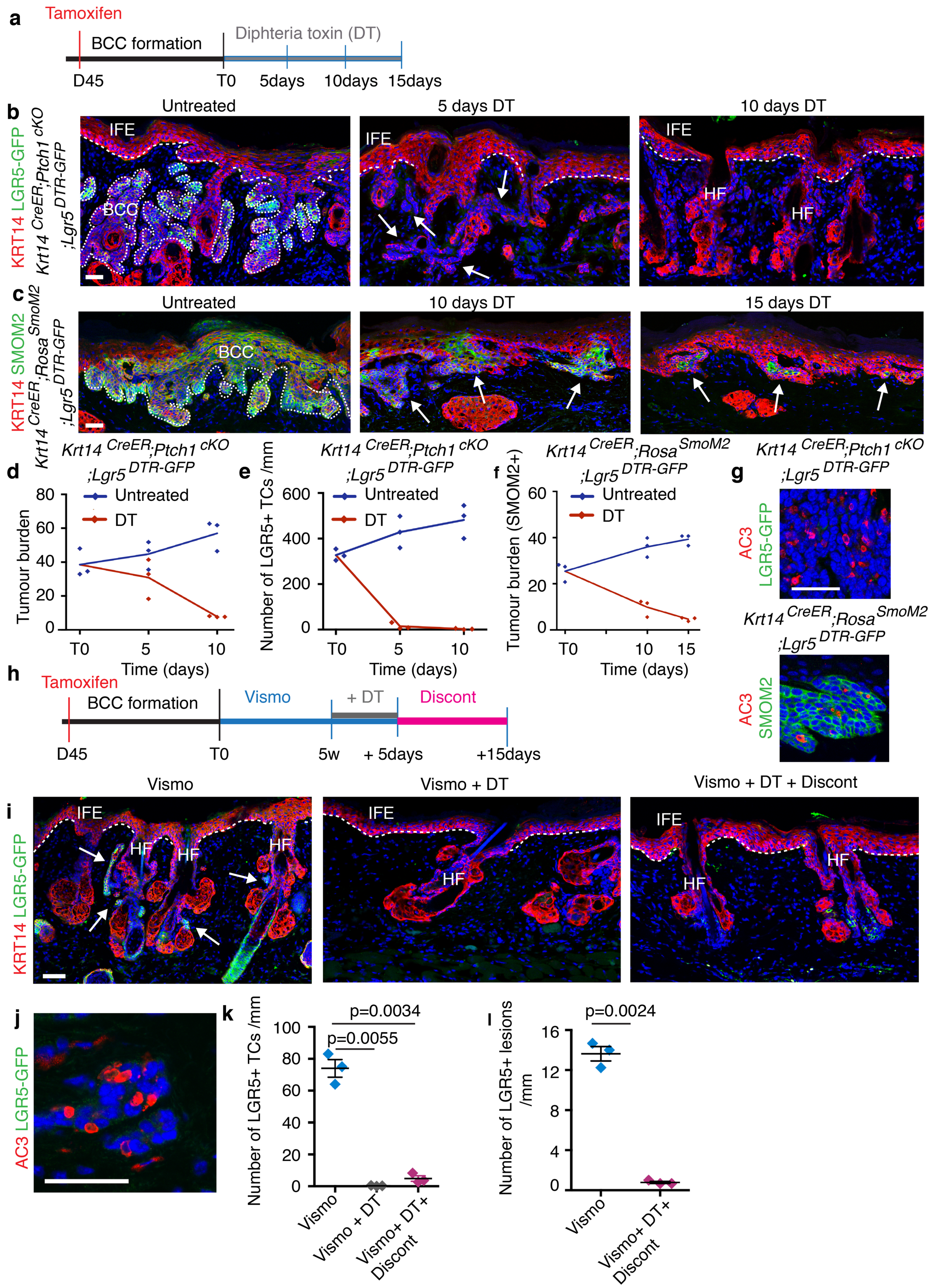
Extended Data Fig. 7 | Vismodegib promotes differentiation of SMOM2-expressing cells during BCC initiation and in *Ptch1*^{cKO} tumour cells. **a**, Protocol for tumour induction and timing of vismodegib administration to *Krt14*^{CreER};*Rosa*^{SmoM2} mice. **b**, Quantification of surviving SMOM2 clones in the interscale (tail epidermis) in untreated mice and after different durations of vismodegib treatment ($n = 3$ mice per time point and condition). Centre values show mean. See Source Data for description of total number of clones counted per time point and condition. **c**, Immunostaining for KRT31 and SMOM2 in whole-mount tail skin (left) and orthogonal views of the clones highlighted in the left panel stained for β4-integrin and SMOM2 (right). **d**, Quantification (mean ± s.e.m.) of the type of SMOM2-expressing clones after different durations of vismodegib treatment ($n = 3$ or 4 mice as indicated in the graph, total number of lesions quantified indicated in parentheses). **e**, **f**, Immunostaining for KRT10 and SMOM2 (**e**) and for LHX2 and SMOM2 (**f**) in untreated and vismodegib-treated mice. Three independent experiments per condition were analysed with similar results. **g**, mRNA expression of genes upregulated in the LGR5$^+$LRIG1$^+$CD71$^+$ population compared to the LGR5$^+$LRIG1$^+$CD71$^-$ population obtained by quantitative PCR ($n = 3$ mice). Bars represent the average fold change over LGR5$^+$LRIG1$^+$CD71$^-$ cells and error bars the s.e.m. **h**, Immunostaining for LGR5–GFP, BrdU and KRT10 in mice that received three injections of BrdU followed by two weeks of vismodegib administration. Three independent experiments per condition were analysed with similar results. Hoechst nuclear staining in blue; scale bars, 100 µm (**c**) and 50 µm (**e**, **f**, **h**).

a

	Sex	Age	Body location	BCC subtype	Treatment duration (months)	Reason for discontinuation	Relapse (months after discontinuation)	Second round treatment
Patient 1	M	57	Shoulder	Superficial	9	Max repsonse	Yes (6)	
Patient 2	M	82	Ear	Nodular and infiltrative	7	Max response & side effects	Yes (9)	
Patient 3	F	84	Nose	Infiltrative	9	Max response & side effects	No (4.5 mo since discont)	
Patient 4	M	80	Forehead	Nodular and infiltrative	13	Max response & side effects	Yes (11)	Ongoing
			Nose	Infiltrative				
			Eyelid	Nodular				

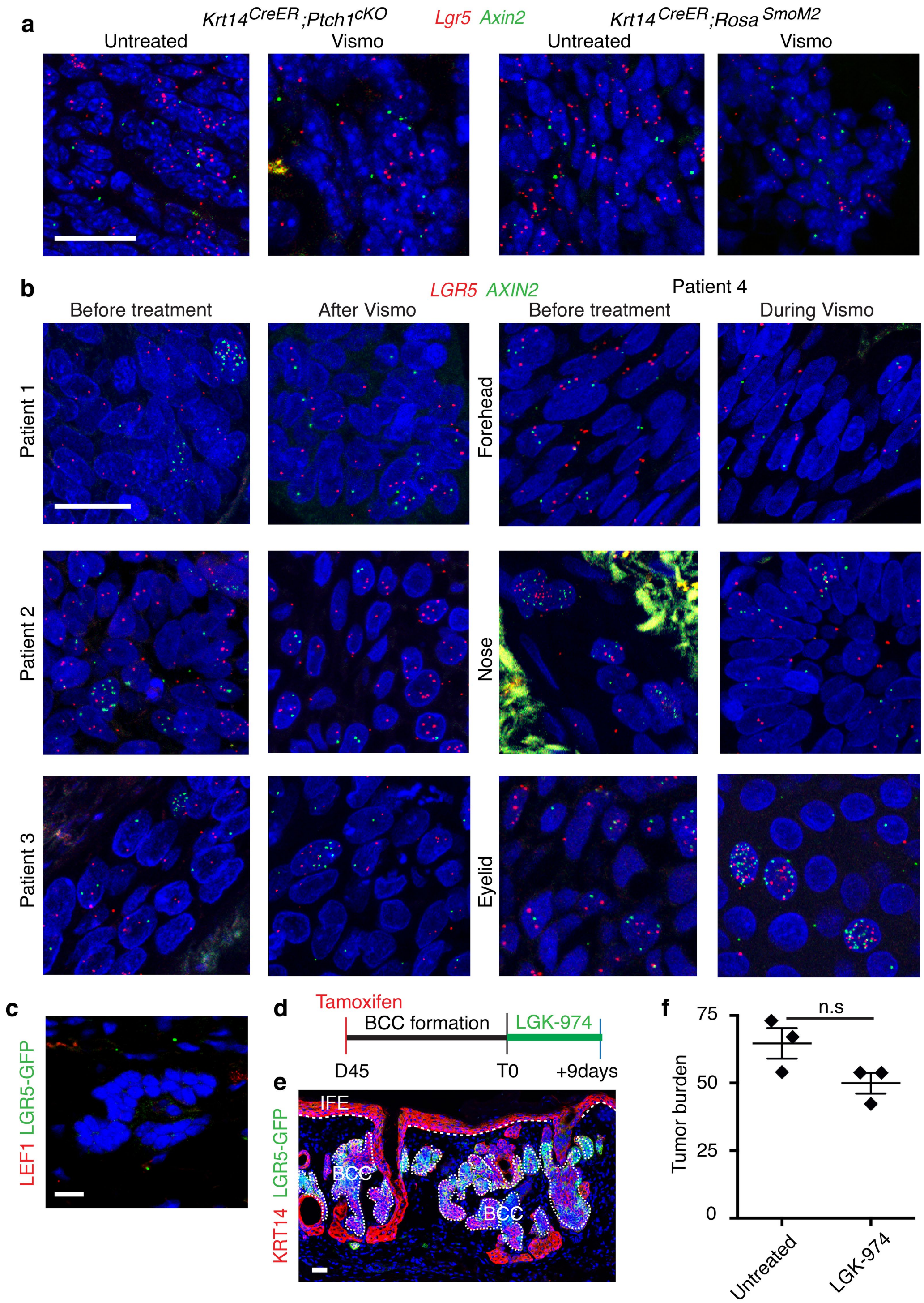


Extended Data Fig. 8 | LGR5 expression in vismodegib-persistent lesions in human BCCs. a, Tables summarizing the BCC and treatment characteristics in the patients analysed. **b**, Immunohistochemistry for KRT14 in biopsies before, after and during vismodegib treatment. **c**, In situ hybridization for *LGR5* and *GLI1* in biopsies from patients before, during and after vismodegib treatment. **d**, Percentage (mean ± s.e.m.) of tumour cells (LGR5$^+$ and LGR5$^-$) that express *GLI1* in biopsies from patients, during or after vismodegib treatment ($n = 3$ samples from different patients (Patients 1–3) or body locations (Patient 4), total number of cells analysed indicated in parentheses). **e**, **f**, Immunohistochemistry for Ki67 (**e**) and KRT10 (**f**) in biopsies before, during and after vismodegib treatment. Hoechst nuclear staining in blue; scale bars, 25 μm.

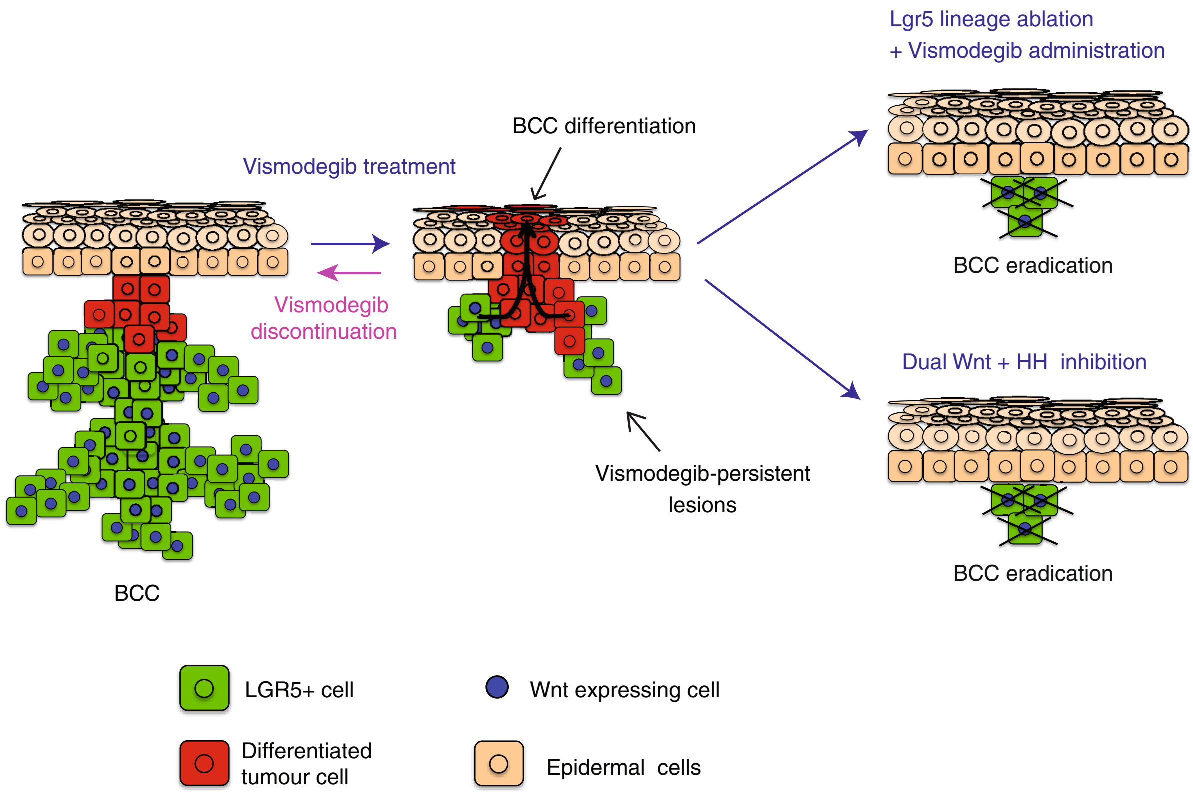


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | *Lgr5* lineage ablation leads to BCC shrinkage and elimination of vismodegib-persistent lesions. a, Protocol for tamoxifen and diphtheria toxin (DT) administration. **b**, **c**, Immunostaining for KRT14 and LGR5–GFP in the $Ptch1^{cKO}$ model (**b**) and for KRT14 and SMOM2 in the *SmoM2* model (**c**) after different durations of DT administration. **d**, Quantification of tumour burden in untreated mice and after DT administration ($n = 3$ mice per time point and condition). Centre values define the mean. See Source Data for description of the skin length and tumour area analysed per mouse. **e**, Number of LGR5–GFP$^+$ tumour cells in untreated conditions and following DT administration ($n = 3$ mice per time point and condition, 1 mm of skin analysed per mouse). Centre values define the mean. **f**, Quantification of tumour burden (SMOM2-expressing cells) in untreated conditions and following DT treatment ($n = 3$ mice per time point and condition). Centre values define the mean. See Source Data for description of the skin length and tumour area analysed per mouse. **g**, Immunostaining for active caspase-3 (AC3) and LGR5–GFP (top) and for active caspase-3 and SMOM2 (bottom) after five administrations of DT. Three independent experiments per condition were analysed with similar results. **h**, Experimental strategy for combination of vismodegib treatment and *Lgr5* ablation in $Krt14^{CreER}$;$Ptch1^{cKO}$;$Lgr5^{DTR-GFP}$ mice. **i**, Immunostaining for LGR5–GFP and KRT14 in $Krt14^{CreER}$;$Ptch1^{cKO}$;$Lgr5^{DTR-GFP}$ mice upon treatment, *Lgr5* ablation and discontinuation. **j**, Immunostaining for active caspase-3 and LGR5–GFP following administration of vismodegib and DT. **k**, Quantification of the number of LGR5–GFP$^+$ cells (mean $\pm$ s.e.m.) in the different experimental conditions upon treatment and discontinuation ($n = 3$ mice, 3 mm of skin analysed per mouse). Two-sided *t*-test. **l**, Quantification of the number of LGR5$^+$ lesions (mean $\pm$ s.e.m.) per length of epidermis (mm) in mice treated with vismodegib and upon discontinuation of treatment with vismodegib and DT ($n = 3$ mice, 3 mm of skin analysed per mouse). Two-sided *t*-test. Hoechst nuclear staining in blue; scale bars, 50 μm. Dashed line delineates basal lamina separating IFE from the dermis. Dotted line delineates BCC. Arrows indicate tumorigenic lesions in **b**, **c** and indicate vismodegib-persistent lesions in **i**.



Extended Data Fig. 10 | Wnt signalling is active in vismodegib-persistent lesions in mouse and human BCCs. **a**, ISH for *Lgr5* and *Axin2* in untreated and vismodegib-treated lesions from *Ptch1*^{cKO} and *SmoM2* mice. **b**, ISH for *LGR5* and *AXIN2* in biopsies from patients before, during and after vismodegib treatment. **c**, Immunostaining for LEF1 and LGR5–GFP in *Ptch1*^{cKO}-derived tumorigenic lesion following treatment with vismodegib and LGK-974. **d**, Protocol used for LGK-974 treatment in *Ptch1*^{cKO} mice. **e**, Immunostaining for LGR5–GFP and KRT14 in BCC treated with LGK-974 for 9 days from the *Ptch1*^{cKO} model. **f**, Quantification of the tumour burden (mean ± s.e.m.) in mice treated with LGK-974 for 9 days or untreated ($n = 3$ mice). See Source Data for description of skin length and tumour area analysed per mouse. Two-sided *t*-test. Hoechst nuclear staining in blue; scale bars, 25 µm. Dashed line delineates basal lamina separating IFE from the dermis. Dotted line delineates BCC. Three independent experiments per condition were analysed with similar results (**a**, **c**, **e**) and two technical replicates were performed for each sample with similar results (**b**).



Extended Data Fig. 11 | Model. Vismodegib administration promotes tumour cell differentiation leading to BCC regression. However, upon vismodegib treatment a small proportion of LGR5$^+$ BCC cells persists, forming vismodegib-tolerant lesions that are slow cycling and characterized by Wnt signalling activation. Discontinuation of vismodegib treatment results in proliferation of LGR5-persistent lesions that lead to BCC relapse. Vismodegib treatment in combination with *Lgr5* lineage ablation or Wnt signalling inhibition results in eradication of BCCs.

https://doi.org/10.1038/s41586-018-0548-6

Genomes of Asgard archaea encode profilins that regulate actin

Caner Akıl[1,2] & Robert C. Robinson[1,2,3]*

The origin of the eukaryotic cell is unresolved[1,2]. Metagenomics sequencing has recently identified several potential eukaryotic gene homologues in Asgard archaea[3,4], consistent with the hypothesis that the eukaryotic cell evolved from within the Archaea domain. However, many of these eukaryotic-like sequences are highly divergent and the organisms have yet to be imaged or cultivated, which brings into question the extent to which these archaeal proteins represent functional equivalents of their eukaryotic counterparts. Here we show that Asgard archaea encode functional profilins and thereby establish that this archaeal superphylum has a regulated actin cytoskeleton, one of the hallmarks of the eukaryotic cell[5]. Loki profilin-1, Loki profilin-2 and Odin profilin adopt the typical profilin fold and are able to interact with rabbit actin—an interaction that involves proteins from species that diverged more than 1.2 billion years ago[6]. Biochemical experiments reveal that mammalian actin polymerizes in the presence of Asgard profilins; however, Loki, Odin and Heimdall profilins impede pointed-end elongation. These archaeal profilins also retard the spontaneous nucleation of actin filaments, an effect that is reduced in the presence of phospholipids. Asgard profilins do not interact with polyproline motifs and the profilin–polyproline interaction therefore probably evolved later in the Eukarya lineage. These results suggest that Asgard archaea possess a primordial, polar, profilin-regulated actin system, which may be localized to membranes owing to the sensitivity of Asgard profilins to phospholipids. Because Asgard archaea are also predicted to encode potential eukaryotic-like genes involved in membrane-trafficking and endocytosis[3,4], imaging is now necessary to elucidate whether these organisms are capable of generating eukaryotic-like membrane dynamics that are regulated by actin, such as are observed in eukaryotic cell movement, podosomes and endocytosis.

Recently, metagenomics studies have identified genes from Asgard archaea (Heimdall, Loki, Thor and Odin) that are homologous to eukaryotic genes that encode machineries involved in membrane

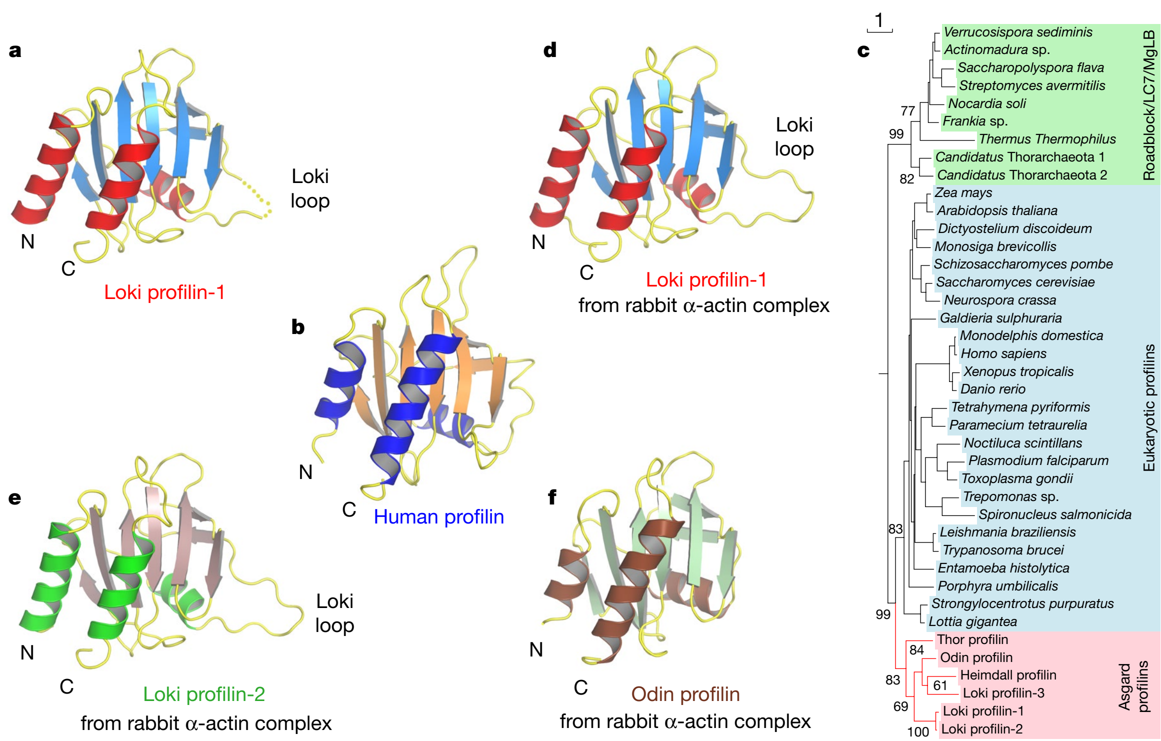


Fig. 1 | Asgard metagenomes encode genuine profilins. a, Schematic of the structure of Loki profilin-1. The partially disordered extended loop (Loki loop) is indicated by a dotted line. N and C indicate the respective termini. Data collection and refinement statistics are found in Supplementary Table 2a. **b**, The structure of human profilin-1 for comparison (RCSB Protein Data Bank (PDB) code: 1FIL). **c**, Profilin phylogenetic tree of Asgard and eukaryotic profilins calculated from structure-based sequence alignment using the Asgard profilin structures in this figure. Protein sequences of the Roadblock/LC7/MgLB group are used as an outgroup because their structures have similar topologies to profilins. Sequence and PDB accession codes used in the alignment are given in Supplementary Table 1b. **d**–**f**, Loki profilin-1 (**d**), Loki profilin-2 (**e**) and Odin profilin (**f**) structures taken from the rabbit α-actin complex structures. Details are found in Fig. 3.

[1]Institute of Molecular and Cell Biology, A*STAR (Agency for Science, Technology and Research), Biopolis, Singapore, Singapore. [2]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [3]Research Institute for Interdisciplinary Science, Okayama University, Okayama, Japan. *e-mail: rrobinson@imcb.a-star.edu.sg

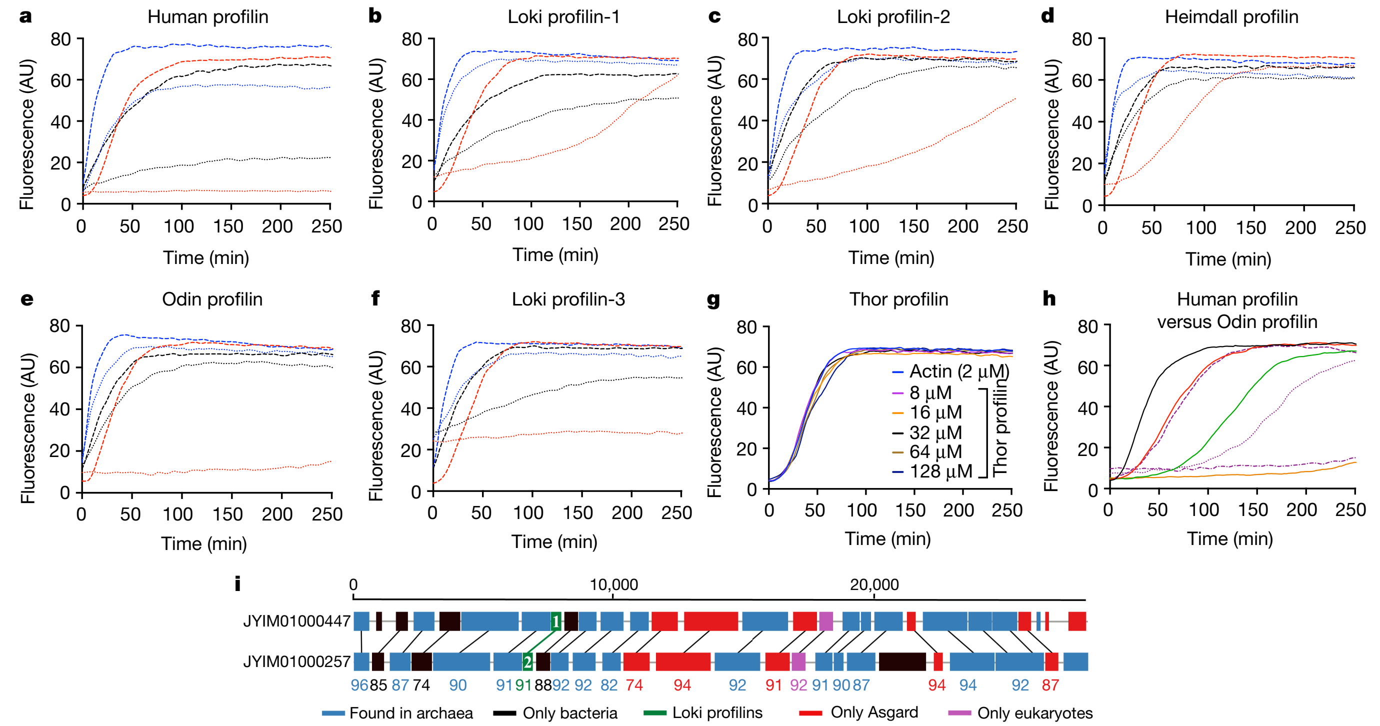


Fig. 2 | Asgard profilins modulate polymerization of mammalian actin in vitro. a, Pyrene–actin polymerization profiles of 2 μM rabbit α-actin (dashes, 10% pyrene–actin) or 2 μM rabbit α-actin with 128 μM human profilin-1 (dots) either initiated alone (red), initiated in the presence of 0.3 μM non-fluorescent actin seeds (blue), or initiated in the presence of 0.3 μM non-fluorescent gelsolin-capped actin seeds (black). **b**–**f**, Polymerization profiles as in **a**, using the specified Asgard profilin at 256 μM (instead of 128 μM human profilin-1). Loki profilin-3 and, to a lesser extent, Loki profilin-1 showed a marked increase in fluorescence on mixing with pyrene–actin; however, their profiles appear to be typical for Asgard profilins, albeit superimposed upon the initial increases. The basis of the increase is unknown, but we speculate that it may be due to oligomer formation. Titrations and expansions of the lag phase regions are in Extended Data Figs. 2, 3. **g**, Pyrene–actin polymerization profiles of 2 μM rabbit α-actin titrated with increasing concentrations of Thor profilin. Thor profilin was not observed to have profilin activity and is not included in subsequent discussions of Asgard profilins. **h**, Comparison of the inhibition of actin nucleation in the pyrene–actin assay reveals that human profilin-1 (solid lines; red 2 μM, green 4 μM, orange 32 μM) is approximately eightfold more potent than Odin profilin (purple; dashes 16 μM, dots 64 μM, dots-and-dashes 256 μM). Actin control (2 μM) is shown as a solid black line. Comparisons for other Asgard profilins are shown in Extended Data Fig. 3b. **i**, Schematic alignment of the parent contigs that contain the Loki profilin-1 and Loki profilin-2 genes[4]. Lines connect homologous genes, with nucleotide percentage identities indicated below (Supplementary Table 3b). Genes with homologues previously found in archaea are coloured blue, and genes currently unique to Asgard archaea are coloured red. Genes with homologues that have only been found to date in bacteria or eukaryotes are coloured black or purple, respectively. The Loki profilins are in green. AU, arbitrary units.

maintenance and function, including trafficking, *N*-glycosylation, ribosomes, endosomal sorting complexes required for transport, the ubiquitination system, and cytoskeletal processes that include actin and profilin homologues[3,4]. However, none of the products of these genes has been shown to be functional at the protein level and the possibility of eukaryotic contamination in the metagenomes has been debated[7,8].

In eukaryotes, force from directed actin polymerization drives membrane remodelling through filament nucleation and elongation machineries—such as formins, the ARP2–ARP3 (also known as ARP2/3) complex and vasodilator-stimulated protein (VASP)—that recruit the profilin–actin complex to initiate and/or sustain polymerization[9]. Asgard archaea possess eukaryotic-like actin genes that encode proteins that display 58–60% identity at the protein-sequence level to human cytoplasmic β-actin and 52–62% identity to actins in other eukaryotes. However, the Asgard actin genes encode products that exhibit a higher level of sequence identity (67–78%) among themselves (Supplementary Table 1a). Phylogenetic analysis revealed that Asgard actins form a separate clade from eukaryotic actins, which indicates that they are related to—yet distinct from—the eukaryotic actins[3,4] (Extended Data Fig. 1a). Structural modelling of Asgard actins yields high-confidence models that are notably similar to eukaryotic actins (Extended Data Fig. 1b). Thus, Asgard actins are both distinct and highly conserved, and are likely to have eukaryotic-like properties. In contrast to Asgard actins, Asgard profilin-like proteins display a low level of sequence identity (11–17%) to human profilin-1 and other eukaryotic profilins (7–24%) (Supplementary Table 1b). This raises questions about their authenticity. To address this issue, we further explored the properties of Asgard profilin-like proteins.

To verify that Asgard archaea possess genuine profilins, we determined the X-ray crystal structure of Loki profilin-1 (Fig. 1a). The structure has a similar topology to that of human profilin (Fig. 1b), albeit with different lengths of helices and loops, including one loop that is notably extended (the 'Loki loop'). Structural comparisons reveal that Loki profilin-1 is divergent from all elucidated eukaryotic profilins (root mean square deviation (r.m.s.d.) > 2.3 Å, Supplementary Table 3a). By contrast, human profilin-1 shows a range of structural divergence, being structurally more similar to profilins from species with which humans share a recent common ancestor (human-to-mouse profilin r.m.s.d = 1.3 Å, Supplementary Table 3a). Phylogenetic analysis (Fig. 1c) using structure-based sequence alignments shows that Asgard profilins also form a separate clade that is distinct from eukaryotic profilins, which indicates that they are related to—yet different from—eukaryotic profilins.

Our attempts to produce Asgard actins in heterologous host expression systems failed to yield appropriate quantities of pure protein for biochemical studies. However, owing to the homology between Asgard and eukaryotic actins, we reasoned that Loki profilin-1 might interact with mammalian actin in vitro. We used the fluorescence of pyrene–actin to follow the time course of rabbit α-actin polymerization. Two types of actin filament seeds were used to test the effects

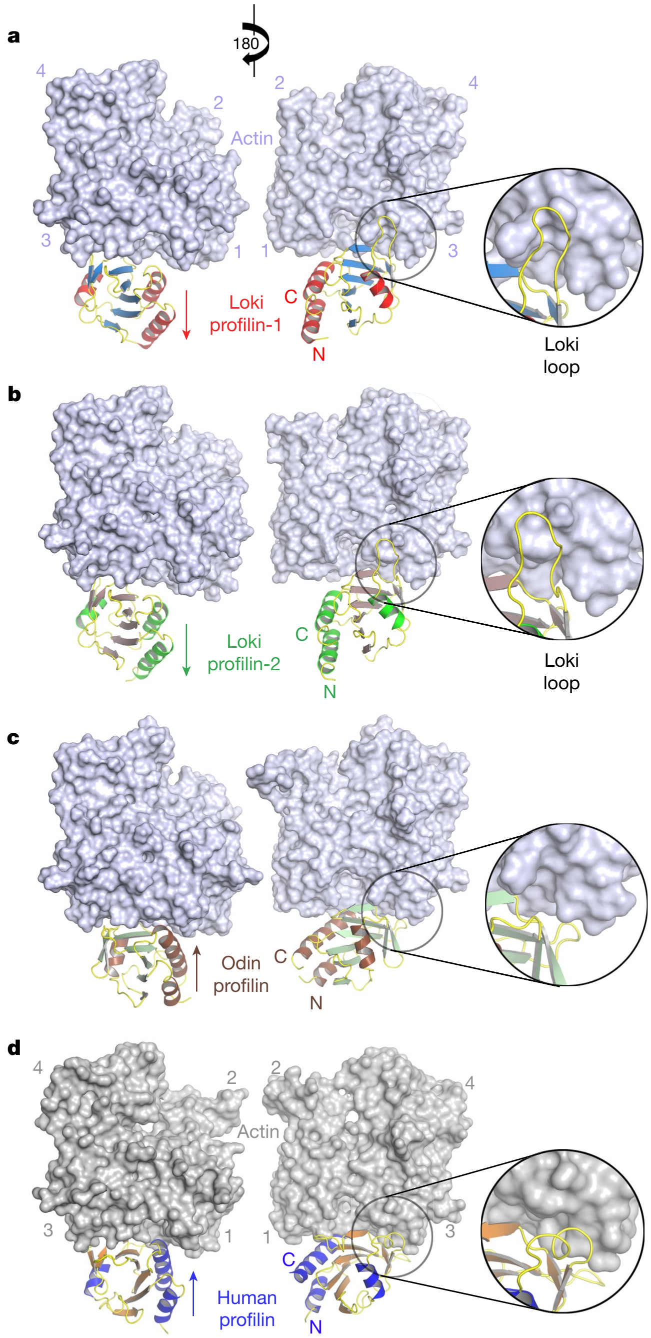


Fig. 3 | The structures of the Loki profilin-1, Loki profilin-2 and Odin profilin complexes with rabbit α-actin. a, Back and front views of the structure of the Loki profilin-1–rabbit α-actin complex. Rabbit α-actin is shown as a surface and Loki profilin-1 is shown in schematic representation. **b**, The Loki profilin-2–rabbit α-actin complex. **c**, The Odin profilin–rabbit α-actin complex. **d**, Structure of the published human profilin-1–rabbit α-actin complex (PDB code: 2PAV) for comparison[12]. Actin subdomains are numbered in **a**, **d**. Arrows indicate relative displacements of the C-terminal helix of each profilin. On the right in each panel, N and C indicate the respective profilin termini. Expanded areas indicate the Loki-loop regions. The actin-interaction residues on the profilins are not conserved between human and Asgard profilins (Extended Data Fig. 5a). Similar observations have previously been noted between profilins from different branches of the eukaryotes, such as between animals and yeast[16]. The complexes were crystallized in the presence of Ca^{2+}, ATP and latrunculin B, a natural product from *Latrunculia* sponges that sequesters G-actin. Relative orientations of the profilin–actin complexes, the latrunculin B-binding sites, comparisons of Asgard profilin folds and examples of electron density are shown in Extended Data Figs. 4, 6, and data collection and refinement statistics are found in Supplementary Table 2b.

of Loki profilin-1 on the two ends of actin filaments. Low concentrations of human profilin-1 or Loki profilin-1 did not slow elongation at the barbed ends (Extended Data Figs. 2, 3). Higher concentrations of human profilin-1 and Loki profilin-1 showed small decreases in elongation rates (Fig. 2a, b), consistent with dynamic barbed end binding by profilin[10], but at least tenfold more archaeal profilin than human profilin was required to achieve similar effects (Extended Data Figs. 2, 3). Gelsolin–actin seeds grow only at the pointed ends. As with lower concentrations of human profilin-1, high concentrations of Loki profilin-1 partially inhibited the elongation of pointed ends (Fig. 2a, b). Spontaneous polymerization of actin monomers depends on a slow, rate-limiting nucleation step. Loki profilin-1 slowed the time course of actin polymerization, but required concentrations that were more than 30 times higher than those required when using human profilin-1 (Fig. 2a, b). Because barbed ends elongate under these conditions, this experiment demonstrates that high concentrations of Loki profilin-1 inhibit spontaneous nucleation; presumably, the higher concentrations of Loki profilin-1 that are required are due to its low affinity for rabbit α-actin. These data suggest that, despite the divergence between Lokiarchaeota and eukaryotes, Loki profilin-1 is partially functional in regulating mammalian actin in vitro, which in turn indicates a profilin-regulated actin system in these archaea.

To demonstrate that the Loki profilin-1 activity is not an isolated case that may be due to eukaryotic contamination in the metagenomes, we produced five other potential Asgard profilins for the in vitro assays (Extended Data Fig. 3c). Loki profilin-2 shares 87% and 91% identity with Loki profilin-1 at amino acid and nucleotide levels, respectively. Comparison of their parent contigs reveals global homology and a high percentage of typical archaeal genes (Fig. 2i). Thus, Loki profilin-1 and Loki profilin-2 appear to come from two related strains of Lokiarchaeota. Loki profilin-2 displayed similar activity to Loki profilin-1 in the polymerization assay (Fig. 2c). Heimdall profilin showed a lesser, but measurable, ability to inhibit spontaneous actin nucleation (Fig. 2d), whereas Odin profilin (Fig. 2e) and Loki profilin-3 (Fig. 2f) showed higher activity relative to Loki profilin-1. The presence of the Loki loop in Loki profilin-1 and Loki profilin-2, but not in Loki profilin-3, indicates possible functional divergence (Extended Data Fig. 5a). Thor profilin displayed no measurable ability to inhibit actin nucleation (Fig. 2g). Odin profilin was approximately eightfold less potent in inhibiting spontaneous actin nucleation, relative to human profilin-1 (Fig. 2h). Odin profilin and Heimdall profilin showed very weak abilities to prevent pointed-end elongation in the gelsolin–actin seed assay. These data demonstrate that actin-regulating profilins are present in three branches of Asgard archaea, which were sequenced from samples collected from different geographical locations: the Mid-Atlantic Ridge (Loki and Heimdall) and Yellowstone National Park (Odin)[4].

To better understand the interactions of Asgard profilins with actin, we determined the co-crystal structures of Loki profilin-1, Loki profilin-2 and Odin profilin bound to rabbit α-actin (Fig. 3). Loki profilin-1 and Loki profilin-2 adopt indistinguishable folds and bind to actin in a similar orientation, using analogous surfaces, to human profilin-1 (Fig. 1b, d, e, 3a, b, d). Actin subdomain 3 interacts similarly with both Loki profilins through residues that are largely conserved in Loki actin (Extended Data Fig. 5b). However, the C-terminal helices of the Loki profilins are slightly displaced from the binding site on actin subdomain 1, relative to human profilin-1 (Fig. 3a, b, d). Furthermore, the Loki loops become ordered on actin binding, and reside close to the surface of actin subdomain 3 (Fig. 3a, b). Odin profilin forms the most compact structure (Fig. 1f)—consistent with Odin inhabiting a geothermal environment[11]—that binds to actin in a similar orientation to human profilin-1 (Fig. 3c, d). The common modes of interaction indicate that several Asgard archaea possess functional profilin–actin systems, akin to those of eukaryotes.

Eukaryotic profilin is integrated into actin-filament nucleating and elongating machineries through binding to polyproline motifs, using a series of π–π interactions with aromatic residues that lie between the N- and C-terminal helices of profilin[12] (Fig. 4a). Analysis of Asgard

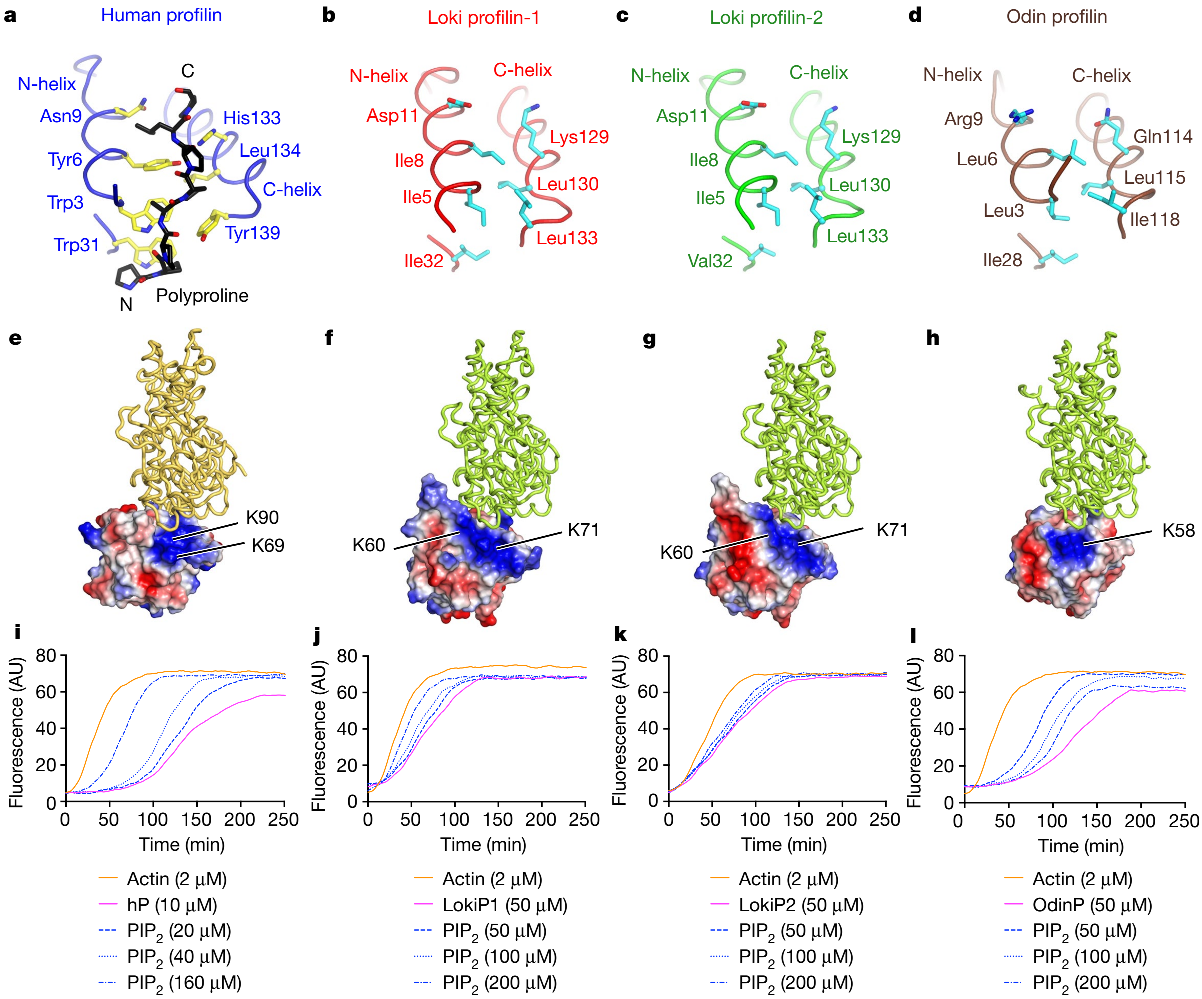


Fig. 4 | Asgard profilins do not bind to polyproline motifs but are sensitive to phospholipids. a, The polyproline-binding site on human profilin-1 (PDB code: 2PAV)[12]. The polyproline ligand is shown in black, with the respective termini labelled as N or C, and the residues on human profilin-1 that interact with the polyproline motif are labelled in blue. **b**–**d**, Equivalent residues to those shown in **a**, from the Loki profilin-1 (**b**), Loki profilin-2 (**c**) and Odin profilin (**d**) structures. In the structures of Asgard profilins, the N- and C-terminal helices are tightly packed relative to those of human profilin-1, and do not display a binding groove. **e**, Positively charged patches on the surface of human profilin-1, with two basic residues indicated. These lie at or near the actin interface (ribbon; PDB code: 2PAV)[12]. **f**–**h**, Similar representations to those shown in **e**, for structures of Loki profilin-1 (**f**), Loki profilin-2 (**g**) and Odin profilin (**h**) complexes with actin, with the basic residues indicated. Views rotated by 180° are shown in Extended Data Fig. 7a. **i**, Pyrene–actin polymerization profiles of rabbit α-actin (2 μM, orange) supplemented with human profilin-1 (hP, 10 μM, pink) and subsequently with increasing concentrations of PtdIns-(4,5)-P_2(1,2-dipalmitoyl) (blue), a soluble version of PIP_2. **j**–**l**, Similar polymerization profiles to those shown in **i**, for Loki profilin-1 (LokiP1, **j**), Loki profilin-2 (LokiP2, **k**) and Odin profilin (OdinP, **l**). These profiles show the diminished effects of the profilins in inhibiting actin polymerization at increasing PIP_2 concentrations. None of the profilins were affected by similar concentrations of inositol trisphosphate (Extended Data Fig. 7b, c).

profilin structures revealed that these residues are absent and that the terminal helices are closer together, eliminating the polyproline-binding groove (Fig. 4a–d). We confirmed using isothermal titration calorimetry that human profilin-1 binds to polyproline sequences, but that Asgard profilins display a $\Delta H = 0$ on mixing with polyproline sequences from VASP (Extended Data Fig. 6). These observations are consistent with the Asgard archaea metagenomes, which do not encode eukaryotic-like polyproline-rich actin nucleation or elongation proteins. Thus, these data further support the claim that Asgard profilins genuinely derive from the Asgard metagenomes and are not the result of eukaryotic contamination[4].

Phosphatidylinositol-4,5-biphosphate (PtdIns(4,5)P_2; also known as PIP_2) is a functional phospholipid for regulating actin at eukaryotic membranes[13]. Although archaea are not known to have the capacity to synthesize PIP_2, they do contain other lipids with inositol phosphate head groups[14]. Surface charge analysis of the human profilin-1–actin and Asgard profilin–actin complexes revealed extensive basic patches that partially overlap with the actin-binding sites that are candidate phospholipid-binding sites (Fig. 4e–h). Thus, we probed whether PIP_2 affects the interaction between Asgard profilins and rabbit α-actin. In the control experiment, increasing concentrations of a soluble version of PIP_2 were able to reverse the inhibition of spontaneous actin nucleation by human profilin-1, seen by a reduction in the profilin-extended lag phase in the pyrene–actin assay (Fig. 4i). This trend was seen for Asgard profilins, although different concentrations of profilin and PIP_2 were needed to suppress the delay in polymerization (Fig. 4j–l). These assays indicate that PIP_2 interacts weakly with Asgard profilins, which mitigates their abilities to prevent actin-filament nucleation. This suggests that Asgard profilin interactions with actin are likely to be regulated by phospholipids; however, the specific phospholipids are unknown.

This structural and biochemical demonstration that Asgard archaea possess primitive phospholipid-sensitive actin-regulating profilins, together with the highly conserved Asgard actin sequences, indicates that Asgard archaea have a functional eukaryotic-like actin machinery. The most parsimonious explanation for its existence is that Asgard archaea and eukaryotes share a common ancestor (with genes for profilin and actin) that arose from within Archaea, as the genomes of other

phyla within this domain do not contain profilins and their actin-like sequences are more divergent from eukaryotic sequences than are those of Asgard archaea. Asgard profilins do not bind to polyproline motifs. Polyproline-directed actin assembly therefore probably appeared later in the eukaryote lineage. Thus, we hypothesize that phospholipid regulation of the profilin–actin complex is a potential mechanism by which force from actin polymerization was integrated into membrane remodelling in the common ancestor of Asgard archaea and eukaryotes. Organisms in all domains of life contain polymerizing filaments constructed from the actin fold. The consensus of known structural and biochemical evidence suggests that each 'actin' subunit binds to a single ATP that is hydrolysed in one cycle of polymerization–depolymerization. Thus, the energy requirements of forming eukaryotic, Asgard or bacterial 'actin' filaments are equal on a subunit-to-subunit comparison basis. However, in many eukaryotic cells the extent and turnover of the actin filament system requires large quantities of ATP[15]. Consequently, imaging of Asgard species is now necessary to understand the complexity of their membrane structures and to determine the extent of their actin cytoskeleton, which will provide insight into the proposed archaea-to-eukaryote transition.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0548-6.

Received: 7 September 2017; Accepted: 23 August 2018;
Published online 3 October 2018.

1. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol*. **15**, 711–723 (2017).
2. López-García, P. & Moreira, D. Open questions on the origin of eukaryotes. *Trends Ecol. Evol*. **30**, 697–708 (2015).
3. Spang, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
4. Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
5. Gunning, P. W., Ghoshdastider, U., Whitaker, S., Popp, D. & Robinson, R. C. The evolution of compositionally and functionally distinct actin filaments. *J. Cell Sci*. **128**, 2009–2019 (2015).
6. Dacks, J. B. et al. The changing view of eukaryogenesis—fossils, cells, lineages and how they all come together. *J. Cell Sci*. **129**, 3695–3703 (2016).
7. Spang, A. et al. Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genet*. **14**, e1007080 (2018).
8. Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A. & Forterre, P. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet*. **13**, e1006810 (2017).
9. Xue, B. & Robinson, R. C. Guardians of the actin monomer. *Eur. J. Cell Biol*. **92**, 316–332 (2013).
10. Courtemanche, N. & Pollard, T. D. Interaction of profilin with the barbed end of actin filaments. *Biochemistry* **52**, 6456–6466 (2013).
11. Thompson, M. J. & Eisenberg, D. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *J. Mol. Biol*. **290**, 595–604 (1999).
12. Ferron, F., Rebowski, G., Lee, S. H. & Dominguez, R. Structural basis for the recruitment of profilin–actin complexes during filament elongation by Ena/VASP. *EMBO J*. **26**, 4597–4606 (2007).
13. Senju, Y. et al. Mechanistic principles underlying regulation of the actin cytoskeleton by phosphoinositides. *Proc. Natl Acad. Sci. USA* **114**, E8977–E8986 (2017).
14. Michell, R. H. Inositol lipids: from an archaeal origin to phosphatidylinositol 3,5-bisphosphate faults in human disease. *FEBS J*. **280**, 6281–6294 (2013).
15. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467**, 929–934 (2010).
16. Ezezika, O. C. et al. Incompatibility with formin Cdc12p prevents human profilin from substituting for fission yeast profilin: insights from crystal structures of fission yeast profilin. *J. Biol. Chem*. **284**, 2088–2097 (2009).

Acknowledgements We thank A*STAR for support; W. Burkholder for reagents; A. Kaya, M. Magnitov and E. Balıkçı for technical advice; and B. Venkatesh for valuable discussions. We thank the experimental facility and the technical services provided by: The Synchrotron Radiation Protein Crystallography Facility of the National Core Facility Program for Biotechnology, Ministry of Science and Technology and the National Synchrotron Radiation Research Center, a national user facility supported by the Ministry of Science and Technology, Taiwan; and the Australian Synchrotron, part of ANSTO.

Reviewer information *Nature* thanks T. Ettema, T. Pollard and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.A. and R.C.R. conceived experiments, analysed data and wrote the paper. C.A. performed experiments.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0548-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0548-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to R.C.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Protein expression and purification. The Asgard profilin gene sequences were codon-optimized (*Escherichia coli*), synthesized (GenScript), placed in the pSY5 vector[17]—which includes an N-terminal HRV 3C protease cleavage site (LeuGluValLeuPheGln↓GlyPro) and 8-histidine tag—and transformed into *E. coli* (DE3). The proteins were expressed overnight (18 °C) in terrific broth (TB) medium supplemented with 0.4% glycerol, following induction by 0.2 mM IPTG (isopropyl-D-1-thiogalactopyranodside) at a cell density characterized by an optical density at 600 nm (OD_{600}) of 0.8–0.9. The resultant cell pellets (10 g) were resuspended in binding buffer 50 ml (20 mM HEPES, 500 mM NaCl, 20 mM imidazole and 1 mM TCEP, pH 7.7) supplemented with Triton X-100 (0.01%), protease inhibitor cocktail (Set III, EDTA-free, Calbiochem) and benzonase (2 μl of 10,000 U/μl, Merck). Cell lysis was performed using an ultrasonic cell disrupter Vibra-Cell (Sonics). The proteins bound to a Ni-NTA affinity chromatography column (HisTrap FF GE Healthcare), washed with binding buffer and eluted via cleavage with HRV 3C protease in binding buffer. After HRV 3C protease cleavage, two residues from the tag (Gly-Pro) remained at the N-terminus of the profilins. Subsequently, the profilins were desalted into appropriate ion-exchange loading buffers. Human profilin-1 and Odin profilin were subjected to cation exchange chromatography (HiTrap SP HP, GE Healthcare, loading buffer 20 mM MES, pH 6) and the remaining profilins to anion exchange chromatography (HiTrap Q HP, GE Healthcare, loading buffer 20 mM HEPES, pH 8). Proteins were eluted by a 50-ml gradient of 0–1 M NaCl in their respective buffers. Size-exclusion chromatography (16/60 Superdex 75 PG, GE Healthcare) in the gel filtration buffer (20 mM HEPES, pH 7.5, 150 mM NaCl, 1 mM TCEP) was used as the final purification step for all profilins. Profilin fractions were pooled and concentrated (2000 MWCO Vivaspin concentrator, Vivascience).

The single polyproline motif from VASP (PPPAPPLPAAQ) was expressed from the pSY7 plasmid, which encodes maltose-binding protein (MBP) between the His-tag and protease cleavage sites of pSY5. The MBP–polyproline construct was purified as described for the profilins, with the exception that the protein was not cleaved but was instead eluted from the Ni-NTA affinity column by 500 mM imidazole followed by gel filtration. VASP and human profilin-1 were purified as previously described[18,19]. Freshly prepared rabbit α-actin, purified including a final gel filtration step as previously described[20], was used in the polymerization experiments. Rabbit α-actin was labelled with pyrene as previously described[20].

Pyrene–actin assays. Pyrene–actin polymerization assays were performed with 2 μM rabbit skeletal-muscle G-actin (10% pyrene labelled). To pre-exchange the calcium ion for magnesium, G-actin in buffer A (2 mM Tris, pH 7.4, 0.2 mM ATP, 0.5 mM DTT, 0.2 mM $CaCl_2$, 1 mM Na azide) was incubated with a 20-fold dilution of 20× Mg-exchange buffer (1 mM $MgCl_2$, 4 mM EGTA) for 2 min. Then, actin polymerization was initiated by the addition 10 μl of 10× KMEI actin polymerization buffer (500 mM KCl, 10 mM $MgCl_2$, 10 mM EGTA, 100 mM imidazole-HCl, pH 7.4) in a total volume of 100 μl. All reactions were carried out in 96-well, black, flat-bottomed plates (Corning, Nunc). The fluorescence intensities were monitored at wavelength 407 nm, after excitation at 365 nm with a Safire^2 fluorimeter (Tecan). To test the effect of actin filament seeds in the pyrene–actin assay, actin filament seeds were prepared by polymerizing 2 μM G-actin for 2 h at room temperature, after which the filaments were sheared by continuously passing the solution through a needle (0.77-mm diameter) for one minute. Subsequently, 15 μl of the actin-seed stock was added simultaneously with the KMEI in the pyrene assay. To test the effect of gelsolin-capped actin seeds in the pyrene–actin assay, gelsolin–actin seeds were prepared by mixing 2 μM G-actin with 4 nM gelsolin in buffer A, and polymerized by adding 10× KMI buffer (500 mM KCl, 10 mM $MgCl_2$ and 100 mM imidazole-HCl, pH 7.4) for 2 h at room temperature. Then, 15 μl of this gelsolin-seed stock was used in total volume of 100 μl for the pyrene–actin polymerization assay. Polymerization was initiated by the addition 10 μl of 10× KMI buffer (KMEI buffer without EGTA). To test the effect of phospholipids, different concentrations of a soluble version of PIP_2 phosphatidylinositol-4,5-diphosphate C-16 (PtdIns-(4,5)-P_2(1,2-dipalmitoyl, Cayman Chemicals) and inositol 1,4,5-trisphosphate (Ins(1,4,5)P_3, Echelon Biosciences) were added in the pyrene–actin polymerization assays.

Isothermal titration calorimeter assays. Isothermal titration calorimeter (ITC) experiments were performed using an ITC200 system (MicroCal). Purified recombinant VASP or MBP–polyproline (2 ml, 0.02 mM) was placed in the cell and titrated with 25 injections of 10 μl of ligand (0.2 mM Loki profilin-1 or human profilin-1) with 4 min between each injection. The sample cell was stirred at 245 r.p.m. and the temperature set at 25 °C. In a second experiment, recombinant Loki profilin-1 and human profilin-1 were used at a concentration of 20 μM and titrated with 25 injections of 10 μl each of 200 μM soluble PIP_2 (phosphatidylinositol 4,5-bisphosphate diC16 (PI(4,5)P_2 diC16), Echelon Biosciences). The final molar ratio of ligand to VASP or MBP–polyproline in the ITC experiments was 1.25:1. All protein and ligand pairs were dialysed in the same buffer to prevent buffer mismatch. Ligand–buffer data were subtracted from ligand–reaction data. Data analysis was carried out using Origin 5.0 (MicroCal).

Crystallization. For the Loki profilin-1 crystallization, equal volumes of protein solution (22 mg/ml) and reservoir solution were mixed. Rod-shaped crystals were formed in 100 mM HEPES, pH 8.5 and 30% w/v polyethylene glycol 10,000. Co-crystallization experiments were performed for Loki profilin-1–rabbit α-actin, Loki profilin-2–rabbit α-actin, Odin profilin–rabbit α-actin with latrunculin B (Calbiochem). For actin complex crystallization trials, 400 μM Asgard profilin, 400 μM rabbit α-actin and 4 mM latrunculin B (1:1:10 at molar ratio) were mixed in buffer A. For the Loki profilin-1–rabbit α-actin–latrunculin B complex, rod-shaped crystals formed in 100 mM HEPES, pH 7.0, 20% w/v polyethylene glycol 6000, 200 mM NaCl, 10 mM ATP disodium salt. For the Loki profilin-2–rabbit α-actin–latrunculin B complex, plate-shaped crystals formed in 100 mM PCTP, pH 7.0 and 25% w/v polyethylene glycol 1500. For the Odin profilin–rabbit α-actin–latrunculin B complex, block-shaped crystals formed in 100 mM citrate, pH 5.0 and 20% w/v polyethylene glycol 6000. All crystallization trials were performed at 18 °C using the sitting-drop vapour-diffusion method and crystals were flash-frozen in the crystallization buffer prior to X-ray data collection.

Structure determination, model building and refinement. The structure of Loki profilin-1 was determined from a platinum derivative of a $P2_12_12_1$ Loki profilin-1 single crystal. The derivative was prepared by incubating (12 h) a speck of solid $K_2Pt(NO_2)_4$ in the crystallization drop after the crystal had formed. Peak ($\lambda = 1.0781$ Å) and remote ($\lambda = 1.0507$ Å) anomalous diffraction datasets were collected on an ADSC QUANTUM 210r CCD detector on beamline MX1 (Australian Synchrotron) at 100 K controlled by custom software. Data processing and scaling were performed in XDS (version November 2016)[21] and CCP4-7.0 AIMLESS (version 0.5.29)[20] (Supplementary Table 2a). A single heavy atom site was located, and resulting phases calculated, in the PHENIX suite (version 1.13-2998)[22] using AutoSol. An initial model containing 96 residues was built using to 2.25 Å resolution data in PHENIX AutoBuild[22]. A native Loki profilin-1 dataset from a single $P2_1$ frozen crystal (100 K) was collected to 1.6 Å resolution on a RAYONIX MX-300 HS CCD detector on beamline TPS 05A (NSRRC) controlled by BLU-ICE (version 5.1) at $\lambda = 1.0$ Å. Data were indexed, scaled and merged in HKL2000 (version 715)[23] (Supplementary Table 2a). Molecular replacement using the partial model was carried out in the PHENIX suite (version 1.13-2998)[22] Phaser, and the model extended in AutoBuild[22]. Final manual adjustments to the model and refinement were carried out in Coot (version 0.8.9 EL)[24] and CCP4-7.0 Refmac5[25], respectively. The final model consists of residues 1–56 and 61–134. Subsequently, X-ray data were collected on beamline MX1 (Australian Synchrotron) controlled by custom software for single crystals frozen at 100 K of the Loki profilin-1–rabbit α-actin complex ($\lambda = 1.0$ Å) and Loki profilin-2–rabbit α-actin complex ($\lambda = 0.9537$ Å), and on beamline MX2 (Australian Synchrotron) for a single crystal of the Odin profilin–rabbit α-actin complex ($\lambda = 1.0$ Å). Data were indexed, scaled and merged in XDS (version November 2016)[21] and ccp4-7.0 CCP4-7.0 AIMLESS (version 0.5.29)[20] (Supplementary Table 2b). The structures of the complexes were determined by molecular replacement using the Loki profilin-1 structure and native rabbit actin[20] as search models in the PHENIX suite (version 1.13-2998) Phaser[22], and the manual rebuilding and final refinement were carried out in Coot (version 0.8.9 EL)[24] and CCP4-7.0 Refmac5[25], respectively. The final Loki profilin-1–rabbit α-actin model consists of Loki profilin-1 residues 1–134 and rabbit α-actin residues 5–39 and 52–375. The final Loki profilin-2–rabbit α-actin models (2 in the asymmetric unit) consist of Loki profilin-2 residues 1–134 (chain A and C) and rabbit α-actin residues 5–41 and 49–375 (chain B), and rabbit α-actin residues 5–40 and 49–375 (chain D). Chain C has N-terminal ordered residues Gly-Pro from the purification tag. The final Odin profilin–rabbit α-actin model consists of Odin profilin residues 1–119 and rabbit α-actin residues 3–375. All final models were verified for good stereochemistry in the PHENIX suite (version 1.13-2998)[22] MolProbity[26] (Supplementary Table 2b).

Sequence analyses. Actin and profilin protein sequences were aligned based on structure in PROMALS3D[27] using multiple templates (Supplementary Table 1). Phylogenetic analyses were carried out in Phylogeny.fr (http://www.phylogeny.fr) and displayed in EvolView (http://www.evolgenius.info/evolview). Sequence alignment figures were created using BoxShade (version 3.21) (http://embnet.vital-it.ch/software/BOX_form.html). The Asgard actins were modelled using the I-TASSER server (https://zhanglab.ccmb.med.umich.edu/I-TASSER/)[28]. Analysis of contigs JYIM01000447 and JYIM01000257 was carried out in the Biostrings (version 2.44.2) R package[29]. Sequences and gene intervals were imported using 'seqinr'; pairwise alignments were conducted for all gene pairs between the two contigs at nucleotide level using 'pairwiseAlignment'; and per cent sequence identities were calculated using 'pid'. BLASTP (https://blast.ncbi.nlm.nih.gov/) and InterPro (version 56.0)[30] were used to find protein homologues.

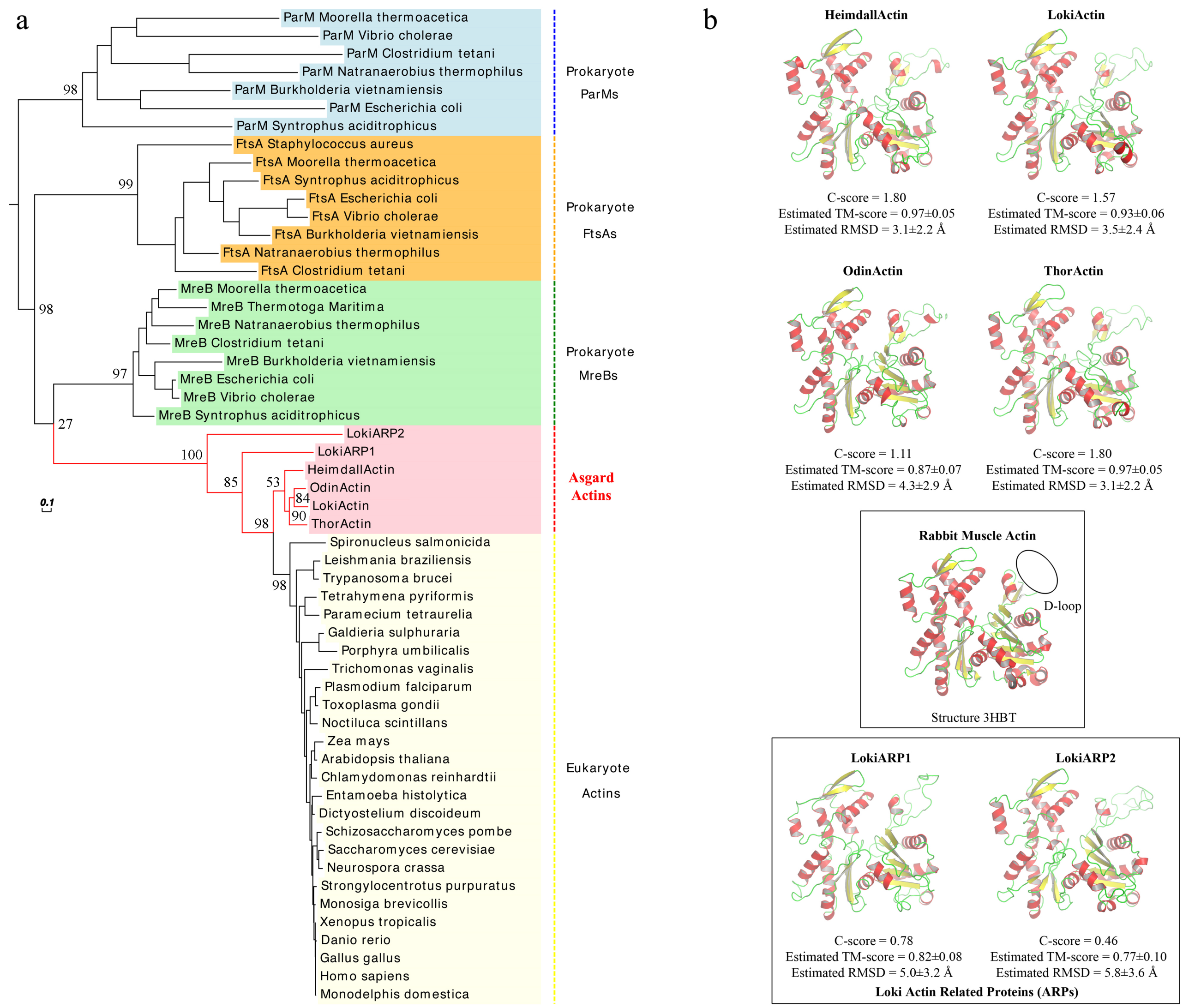
Statistics and reproducibility. ITC experiments were repeated 2 times, and all other biochemical experiments were repeated 3 times, with similar results.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

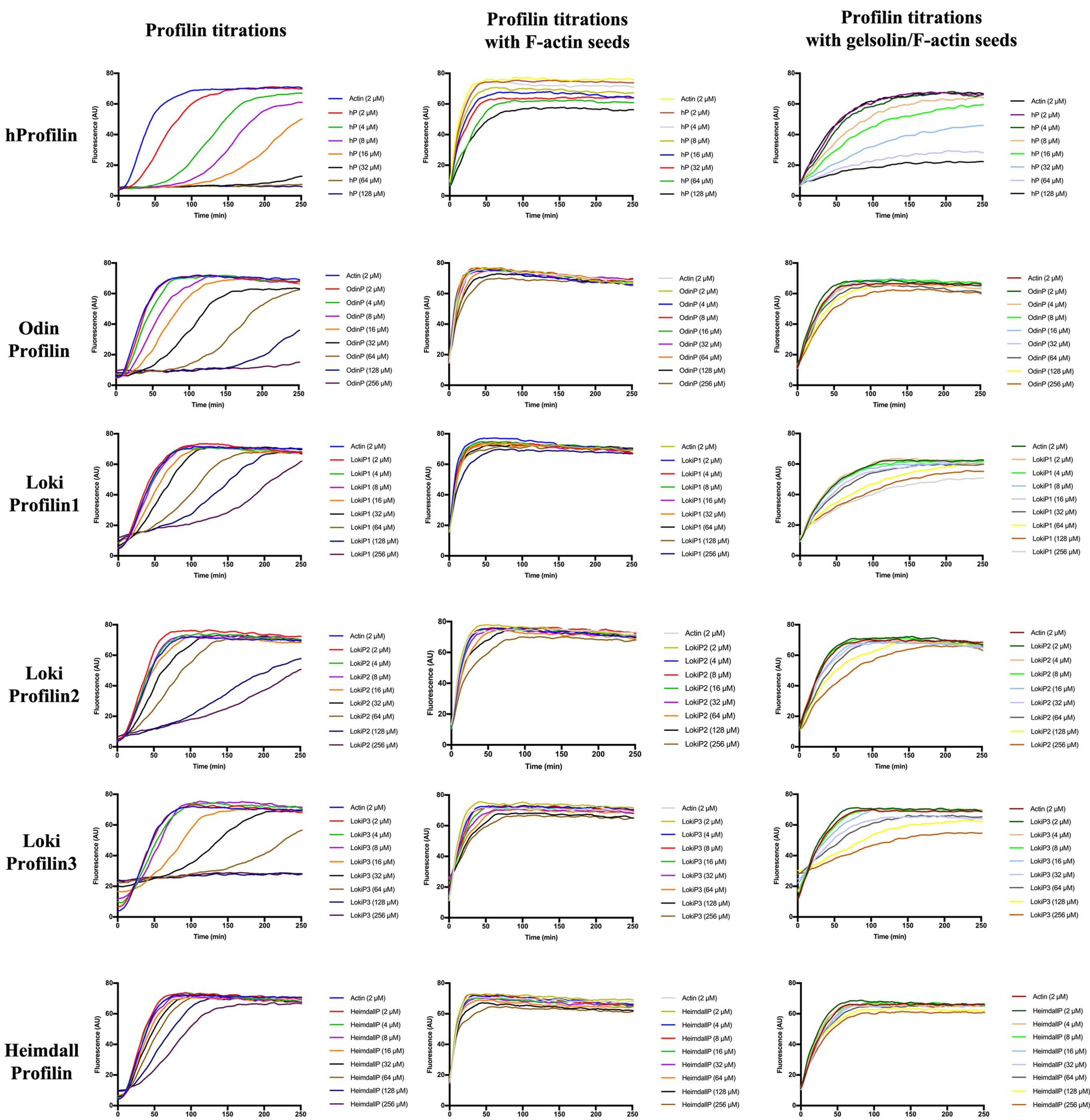
Data availability

The atomic coordinates and structure factors have been deposited in the Protein Data Bank (PDB) under the accession codes 5YED, 5YEE, 5ZZA and 5ZZB. All other data are available from the corresponding author upon reasonable request.

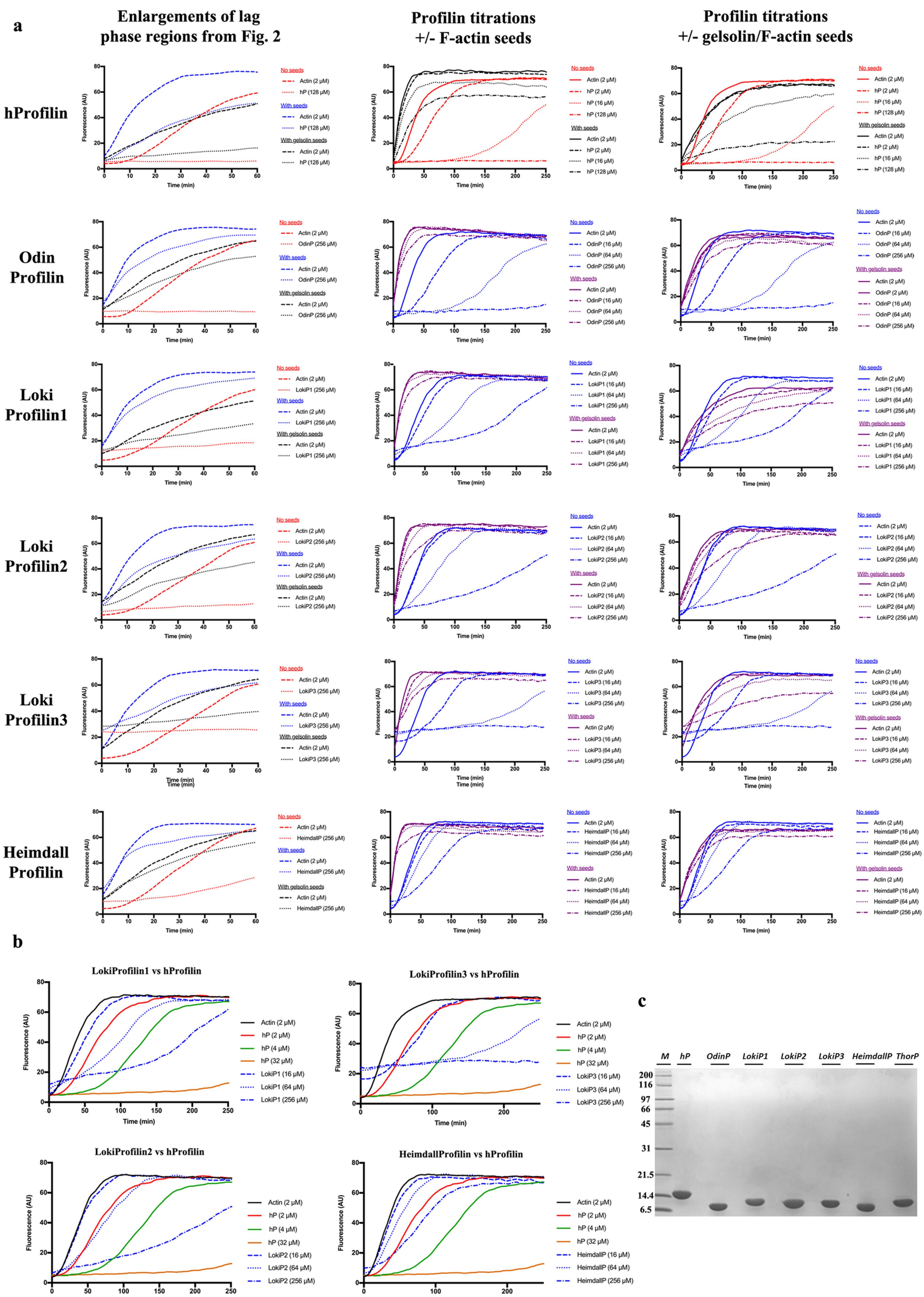
17. Nag, S. et al. Ca^{2+} binding by domain 2 plays a critical role in the activation and stabilization of gelsolin. *Proc. Natl Acad. Sci. USA* **106**, 13713–13718 (2009).
18. Xue, B., Leyrat, C., Grimes, J. M. & Robinson, R. C. Structural basis of thymosin-β4/profilin exchange leading to actin filament polymerization. *Proc. Natl Acad. Sci. USA* **111**, E4596–E4605 (2014).
19. Lee, W. L., Grimes, J. M. & Robinson, R. C. *Yersinia* effector YopO uses actin as bait to phosphorylate proteins that regulate actin polymerization. *Nat. Struct. Mol. Biol*. **22**, 248–255 (2015).
20. Wang, H., Robinson, R. C. & Burtnick, L. D. The structure of native G-actin. *Cytoskeleton (Hoboken)* **67**, 456–465 (2010).
21. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
22. Adams, P. D. et al. The Phenix software for automated determination of macromolecular structures. *Methods* **55**, 94–106 (2011).
23. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol*. **276**, 307–326 (1997).
24. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
25. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
26. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
27. Pei, J. & Grishin, N. V. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol. Biol*. **1079**, 263–271 (2014).
28. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40 (2008).
29. Pagès, H. et al. DebRoyBiostrings: Efficient manipulation of biological strings. https://rdrr.io/bioc/Biostrings/ (2018).
30. Finn, R. D. et al. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res*. **45**, D190–D199 (2017).



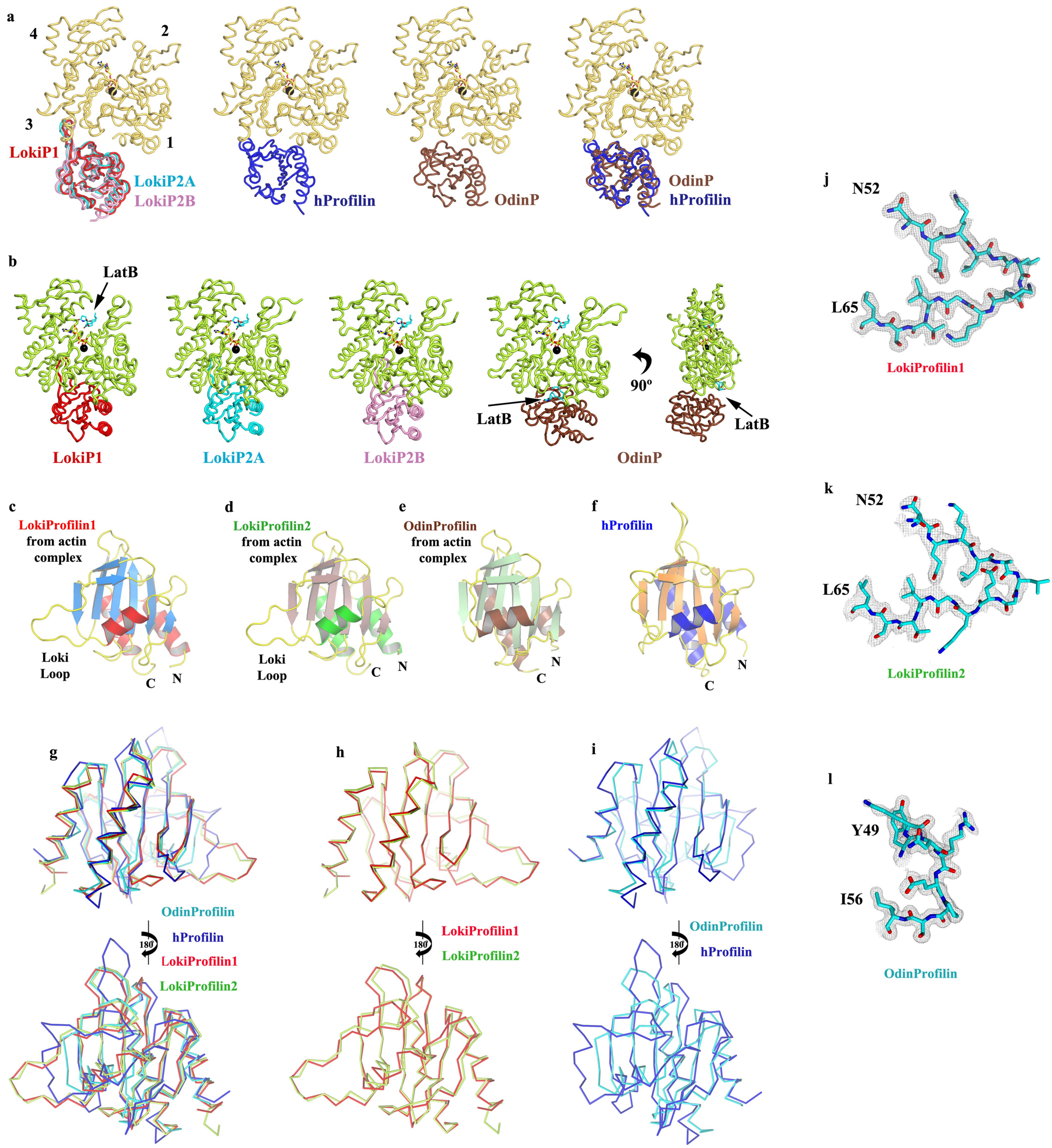
Extended Data Fig. 1 | Asgard actins. a, Phylogenetic tree of the polymerizing actin fold. This phylogenetic tree reveals that the variability observed between the Asgard and eukaryotic actins is approximately similar in magnitude to the variability found within bacterial MreBs and lower than that observed within bacterial FtsAs or ParMs, which indicates a probable conservation in function between the Asgard and eukaryotic actins. Nevertheless, Asgard and eukaryotic actins form two separate clusters, which indicates that the Asgard actin genes are unlikely to be contaminants from eukaryotes. The actin, MreB, FtsA and ParM sequences were aligned in PROMALS3D using accession codes for the actins shown in Supplementary Table 1, and the structures: 1JCE, 3WT0, 1MWM and actin 3HBT (PDB codes). The MreBs are ABO56305.1, ABC18867.1, ACB86382.1, AAC72350.1, ABC76501.1, AAO36573.1 and AAN82446.1; the FtsAs are AAN78610.1, AAO35699.1, ABC76564.1, AAF95541.1, ACB84890.1, ABC19166.1, and KVD45461.1; and the ParMs are ABO60264.1, ABC18679.1, ACB86228.1, BAC79052.1, ABC77529.1, AAO37409.1, and ABV16243.1 (GenBank codes). Heimdall actin, Loki actin, Odin actin and Thor actin are highly similar to eukaryotic actins, as judged by the branch points and identities (Supplementary Table 1a), and are here referred to as the 'Asgard actins'. **b**, Models of the Asgard actins. The Asgard actins were modelled using I-TASSER. The 'C-score' is a confidence score for estimating the quality of predicted models, which is typically in the range of −5 to 2 and in which a high C-score signifies a model with a high confidence. Template modelling scores ('TM-scores') range between 0 and 1, and values greater than 0.5 indicate models of the correct topology and a value of 1 indicates an exact match. The TM-score and r.m.s.d. are estimated by linear regression, and the estimated errors are the root-mean-squared TM-score or r.m.s. deviations. The models of Heimdall actin, Loki actin, Odin actin and Thor actin are of high confidence and are notably similar to the structure of rabbit muscle actin, as judged by these metrics. Models of Loki ARP1 and Loki ARP2 are slightly less reliable and their sequences are more divergent (Supplementary Table 1a). The major differences between the Loki ARP1 and Loki ARP2 models and the structure of rabbit muscle actin lie in the DNAse I binding loop (D-loop), which is often disordered in G-actin structures (indicated by the oval) and becomes ordered on forming F-actin. This region has insertions in Loki ARP1 and Loki ARP2.



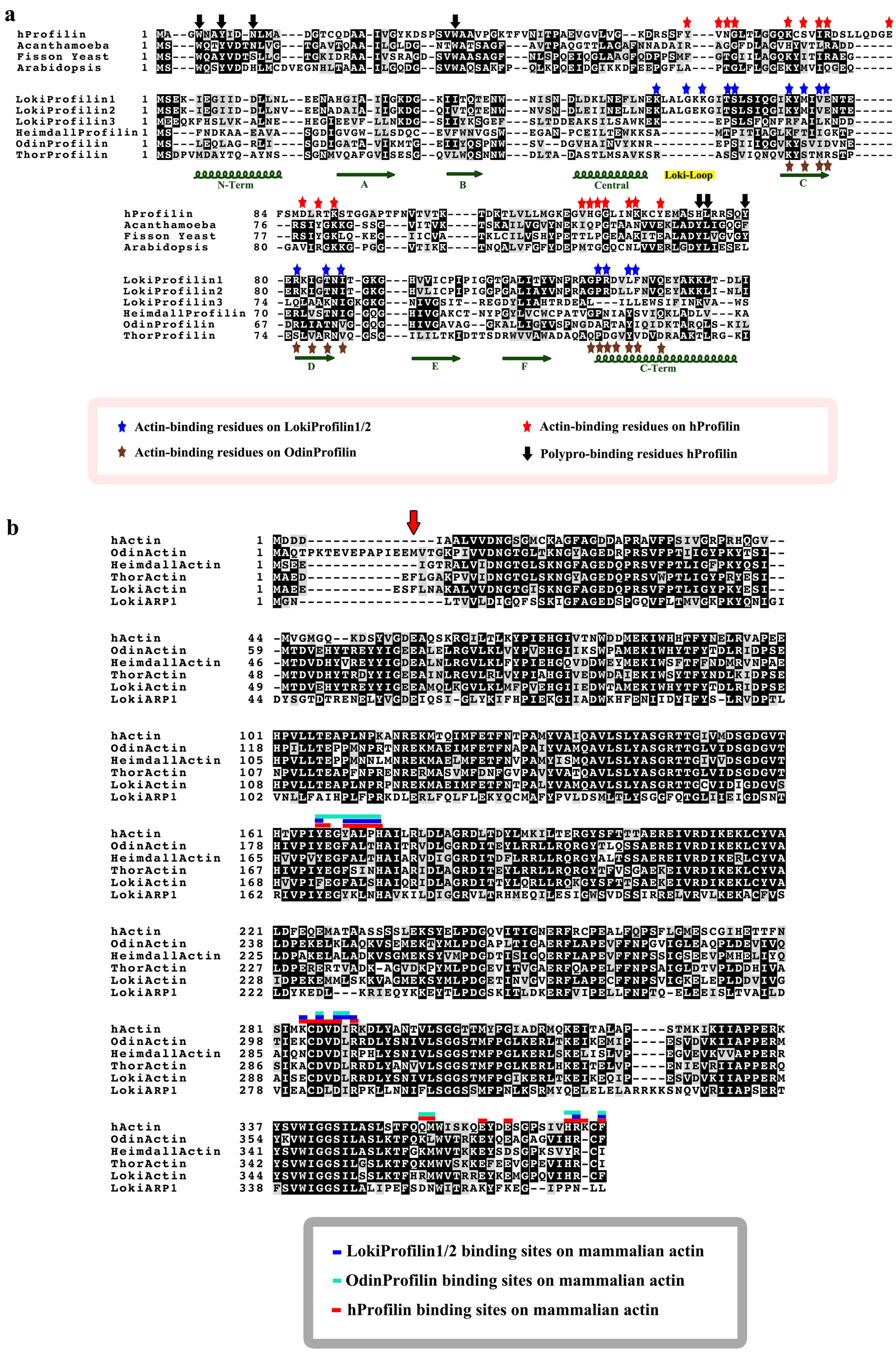
Extended Data Fig. 2 | Titration data. Titration data for the polymerization profiles that are shown in Fig. 2a–f.



Extended Data Fig. 3 | Comparisons of polymerization profiles. Polymerization profiles are shown in Fig. 2a–f. **a**, Enlargement of the lag phases and comparison between non-seeded and seeded actin polymerization are also shown as titrations for each profilin. **b**, Comparison of the inhibition of actin nucleation in the pyrene–actin assay of Asgard profilins, relative to human profilin-1. These profiles complement the Odin actin–human profilin-1 comparison shown in Fig. 2h. **c**, The quality of the purified Asgard profilin proteins. SDS–PAGE gels indicating the purity of the profilins (abbreviated as ‘P’ in this panel) used in this study. M indicates the protein molecular weight (MW) marker lane.

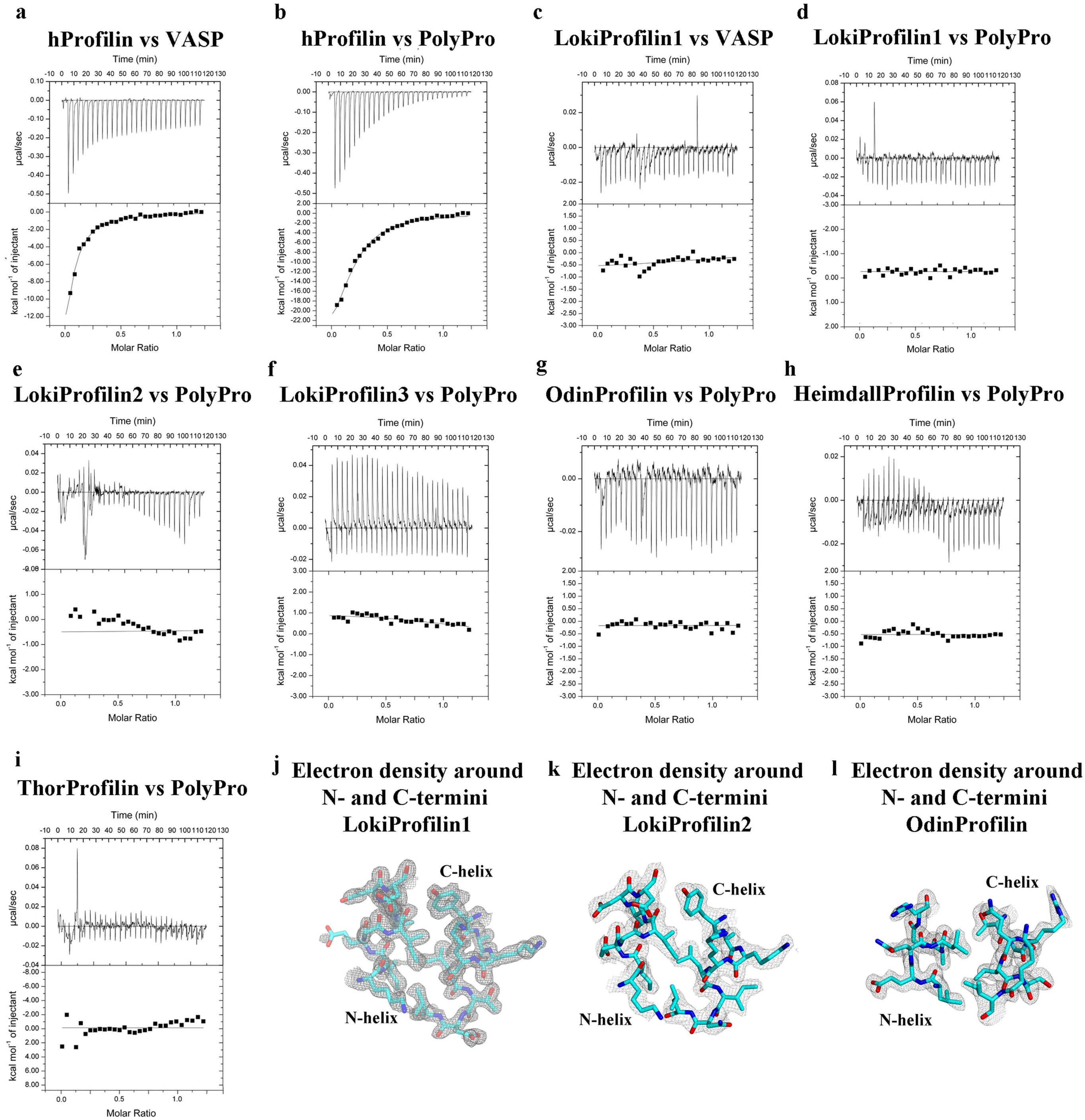


Extended Data Fig. 4 | Comparison of profilin structures. a, Ribbon representations of the profilin–actin complexes. The overlays were created by superimposing the actin structures to show the relative binding geometries of the profilins (abbreviated as P). For clarity, only one actin is displayed. The positions of Loki profilin-1 and Loki profilin-2 overlay well with the major interaction with subdomain 3. The rabbit α-actin subdomains are indicated as black numbers. Some flexibility is observed in the binding mode of Loki profilin-2, as seen by a slight shift between the two complexes in the asymmetric unit (labelled LokiP2A and LokiP2B). Human profilin-1 and Odin profilin adopt binding positions that also interact with subdomain 3, but which additionally have more-intimate contacts with rabbit α-actin subdomain 1. The human profilin-1–rabbit α-actin and Odin profilin–rabbit α-actin complexes are shown separately to aid in interpretation of the overlay. Bound calcium ions are shown as black spheres and ATP in stick representation. **b**, Latrunculin B (LatB, cyan) binds to analogous positions in rabbit α-actin in each complex. The Odin profilin–rabbit α-actin binds to a second molecule of latrunculin B at the edge of the Odin profilin–rabbit α-actin interface. For clarity, two views of Odin profilin–rabbit α-actin are shown. **c**–**f**, Cartoons of the profilin structures rotated by 180° around the *y* axis with respect to Fig. 1b, d, e, f. **g**–**i**, Cα traces of the superimpositions of human, Odin, Loki profilin-1 and Loki profilin-2 structures (**g**), Loki profilin-1 and Loki profilin-2 structures (**h**) and the human and Odin profilins (**i**) to highlight their similarities and differences. **j**–**l**, The $2F_o - F_c$ electron density contoured at 1.0σ—indicating the quality of the structures—is shown surrounding the Loki-loop regions of Loki profilin-1 (**j**) and Loki profilin-2 (**k**), along with the absence of the Loki-loop in this region in Odin profilin (**l**).

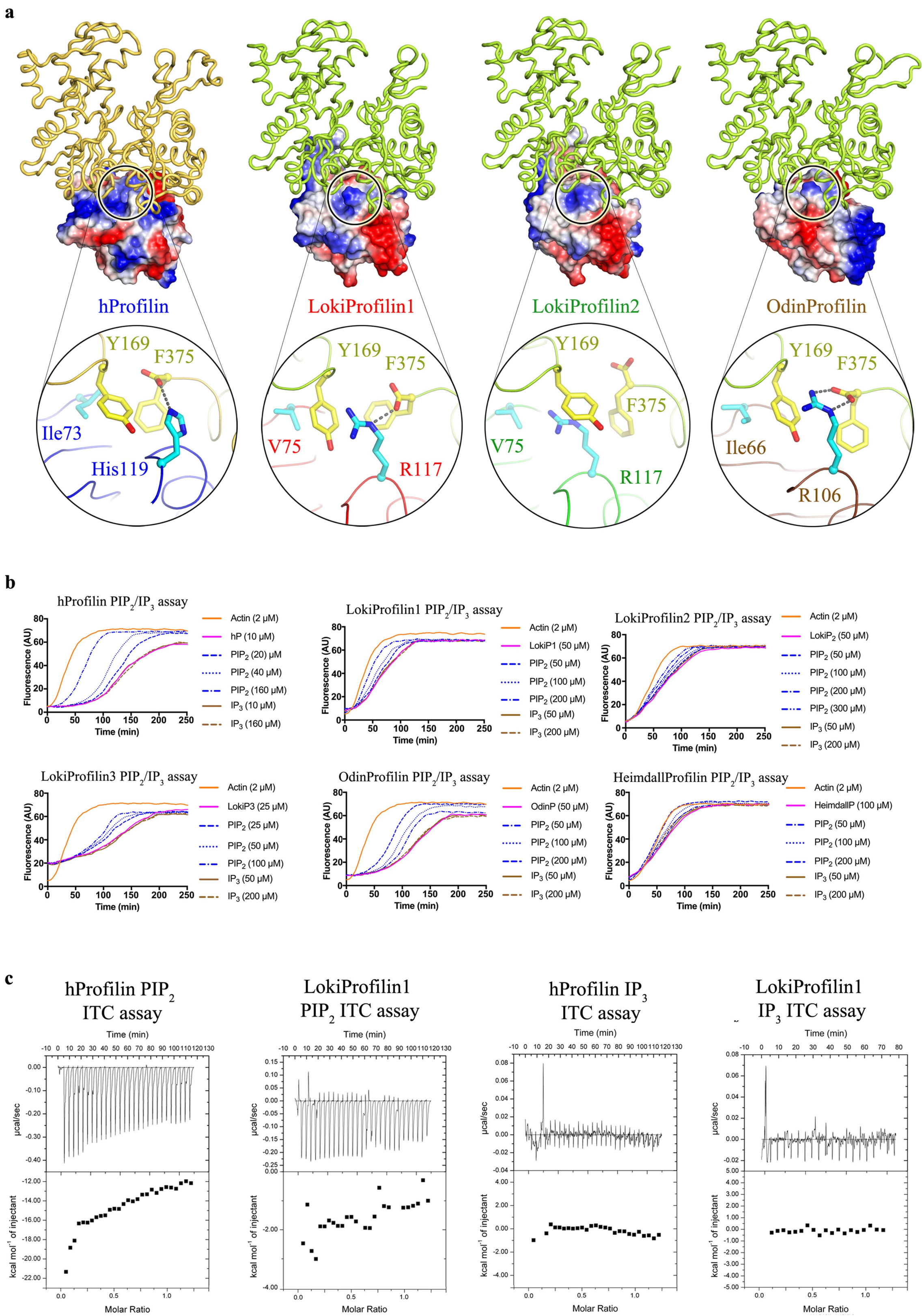


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Structure-based sequence alignments. **a**, Eukaryotic and Asgard profilins. The sequences were aligned using profilin structures (PDB codes: 2PAV, 1ACF, 3D9Y and 1A0K) and Loki and Odin profilin structures as guides. The Loki profilin-1 secondary structure is indicated below the alignment in green. Actin-binding sites are indicated by red stars (human profilin-1), blue stars (Loki profilin-1 and Loki profilin-2) and brown stars (Odin profilin). These appear in the same regions but are not well-conserved between the eukaryotes and archaea in amino acid sequences. Black arrows indicate polyproline-binding resides, which are conserved in the eukaryotic sequences but not in the archaea sequences. **b**, Eukaryotic and Asgard actins. The sequences were aligned using the native rabbit α-actin structure as a guide (PDB code: 3HBT). Profilin-binding sites observed in the structures are indicated by red (human profilin-1; PDB code: 2PAV), blue (Loki profilin-1 and Loki profilin-2) and cyan (Odin profilin) bars above the alignment. The profilins bind to regions of actin that display conservation between human cytoplasmic β-actin and the Asgard actins. Loki ARP1 is included here to demonstrate that other actin-like sequences are found in the Asgard metagenomes, and that these sequences share less sequence identity with human cytoplasmic β-actin. The red arrow indicates an alternative start methionine in the Odin actin sequence, and the Odin actin sequence displayed here takes account of a potential frame-shift in the nucleotide sequence (GenBank accession code: MDVT01000007).



Extended Data Fig. 6 | Human profilin-1 binds to polyproline motifs, whereas Asgard profilins do not. **a**, ITC binding profiles of full length VASP, which contains multiple polyproline motifs, titrated with human profilin-1. **b**, ITC binding profiles of a polyproline motif from VASP (PPPAPPLPAAQ) titrated with human profilin-1. **c**, ITC binding profiles of full-length VASP titrated with Loki profilin-1. **d**, ITC binding profiles of the single polyproline motif from VASP titrated with Loki profilin-1. **e**–**i**, Similar titrations for the single polyproline motif from VASP to the profiles shown in **d**, for Loki profilin-2 (**e**), Loki profilin-3 (**f**), Odin profilin (**g**), Heimdall profilin (**h**) and Thor profilin (**i**) are shown, each of which displays no observable interaction of the Asgard profilins with polyproline sequences. **j**–**l**, The $2F_o-F_c$ electron density, contoured at 1.0σ, surrounding the N- and C-termini of Loki profilin-1 (**j**), Loki profilin-2 (**k**) and Odin profilin (**l**), showing that the terminal helices are tightly packed and well-ordered.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Phospholipid binding. a, Potential phospholipid-binding sites on the surface of Asgard profilins. Actin complexes are shown rotated by 180° around the *y* axis, relative to those shown in Fig. 4e–h. Expanded regions show that the profilins each present a basic residue that binds to the C-terminal residue of actin (Phe375) and to Tyr169, which in turn binds to a conserved Val or Ile on the profilins. Phe375 is conserved, whereas Tyr169 has a conservative substitution (Phe) in Loki actin and Odin actin sequences (Extended Data Fig. 5b). Arg117 and interacting residues on actin (Tyr169 and Phe375) adopt slightly different conformations in the Loki profilin-1–rabbit α-actin and Loki profilin-2–rabbit α-actin complexes. Lys60 (Fig. 4f, g) lies at the start of the Loki loop of Loki profilin-1 and Loki profilin-2. **b**, Pyrene–actin polymerization profiles of rabbit α-actin (2 μM, orange) supplemented with profilins and subsequently with increasing concentrations of a soluble version of PIP_2 or inositol trisphosphate (IP_3). In each case, PIP_2 causes a reduction in the delay of polymerization caused by profilin inhibition of spontaneous actin nucleation, whereas IP_3 has no effect. **c**, ITC was used to verify the interaction between selected profilins and PIP_2. ITC binding profile of human profilin-1 titrated with soluble PIP_2 demonstrated an interaction, whereas Loki profilin-1 titrated with soluble PIP_2 showed marginal binding. Both profilins showed a $\Delta H = 0$ on addition of IP_3 in the ITC assay. The marginal binding observed for Loki profilin-1 with soluble PIP_2 appears to be above the $\Delta H = 0$ observed for IP_3.

https://doi.org/10.1038/s41586-018-0587-z

Handover mechanism of the growing pilus by the bacterial outer-membrane usher FimD

Minge Du[1], Zuanning Yuan[1], Hongjun Yu[1], Nadine Henderson[2,3], Samema Sarowar[4], Gongpu Zhao[5], Glenn T. Werneburg[2,3,6], David G. Thanassi[2,3]* & Huilin Li[1]*

Pathogenic bacteria such as *Escherichia coli* assemble surface structures termed pili, or fimbriae, to mediate binding to host-cell receptors[1]. Type 1 pili are assembled via the conserved chaperone–usher pathway[2–5]. The outer-membrane usher FimD recruits pilus subunits bound by the chaperone FimC via the periplasmic N-terminal domain of the usher. Subunit translocation through the β-barrel channel of the usher occurs at the two C-terminal domains (which we label CTD1 and CTD2) of this protein. How the chaperone–subunit complex bound to the N-terminal domain is handed over to the C-terminal domains, as well as the timing of subunit polymerization into the growing pilus, have previously been unclear. Here we use cryo-electron microscopy to capture a pilus assembly intermediate (FimD–FimC–FimF–FimG–FimH) in a conformation in which FimD is in the process of handing over the chaperone-bound end of the growing pilus to the C-terminal domains. In this structure, FimF has already polymerized with FimG, and the N-terminal domain of FimD swings over to bind CTD2; the N-terminal domain maintains contact with FimC–FimF, while at the same time permitting access to the C-terminal domains. FimD has an intrinsically disordered N-terminal tail that precedes the N-terminal domain. This N-terminal tail folds into a helical motif upon recruiting the FimC-subunit complex, but reorganizes into a loop to bind CTD2 during handover. Because both the N-terminal and C-terminal domains of FimD are bound to the end of the growing pilus, the structure further suggests a mechanism for stabilizing the assembly intermediate to prevent the pilus fibre diffusing away during the incorporation of thousands of subunits.

Type 1 pili are helical structures composed of the major pilus subunit FimA, with a distal tip composed of FimF, FimG and—at the very tip—the FimH adhesin (Fig. 1a, b). FimH binds to mannosylated glycoproteins on the host bladder, thereby enabling the microbe to gain a foothold for infection[6]. Two key concepts have previously been established for pilus biogenesis mediated by the chaperone–usher pathway[3,7]. First, pilus subunits comprise an incomplete immunoglobin-like fold, which lacks the seventh β-strand that is present in canonical immunoglobulin folds. As nascent pilus subunits enter the periplasm via the Sec translocon, the FimC chaperone donates its G1 β-strand to complete the immunoglobulin fold of each subunit and thereby stabilizes each subunit, in a mechanism termed donor-strand complementation[8,9] (Extended Data Fig. 1). Second, pilus subunits polymerize at the FimD usher via a donor-strand exchange mechanism, in which the G1 β-strand of the chaperone is replaced by the N-terminal extension of an incoming pilus subunit[10,11] (Extended Data Fig. 1). Polymerization of the pilus fibre is driven by the folding-energy differential between the β-strand insertions of donor-strand complementation and those of donor-strand exchange.

The FimD usher catalyses ordered pilus biogenesis at the bacterial outer membrane[12] (Fig. 1b). The β-barrel domain of the usher functions as a protein secretion channel and is occluded by a plug domain in the resting (apo) state[13]. Following FimC–FimH recruitment to the N-terminal domain (NTD) of FimD[14,15], the plug exits the channel and moves to the periplasm adjacent to this NTD. The differential affinity of the usher for chaperone–subunit complexes, coupled with the unique capacity of the FimH adhesin to activate the usher, ensures pilus assembly proceeds in an ordered fashion[12,16,17]. The kinetics of donor-strand exchange between pilus subunits also has an important role in determining the order in which subunits are incorporated[18,19]. In the crystal structure of a FimD–FimC–FimH ternary complex, FimC–FimH has dissociated from the NTD of FimD and is bound to the C-terminal domains (CTDs) of FimD[13]. In the crystal structure of a FimD–tip complex (FimD–FimC–FimF–FimG–FimH), the last-incorporated subunit (FimF) in complex with FimC is similarly bound to the CTDs[20]. Clearly, there is a handover event during pilus assembly in which a chaperone subunit that has been recruited to the NTD of the usher transfers to the CTDs, where pilus translocation occurs[13,20,21]. However, the mechanism of this handover and its timing with respect to subunit polymerization—that is, donor-strand exchange of the incoming subunit with the preceding subunit—has previously not been known.

To address these questions, we derived two cryo-electron microscopy (cryo-EM) 3D maps of the FimD–tip complex, one at 4.0-Å resolution (termed conformer 1) and the other at 5.1-Å resolution (termed conformer 2) (Fig. 1c, d, Extended Data Figs. 2, 3, Extended Data Table 1, Supplementary Video 1). Atomic models were built using available component crystal structures (Fig. 2a, b). As expected, both conformers consist of three pilus subunits bound to FimD: the FimH adhesin, which has crossed the usher channel to the extracellular face; FimG, which is inside the FimD chamber; and FimF in complex with the FimC chaperone, on the periplasmic face. The FimD plug domain, NTD, CTD1 and CTD2—all of which are on the periplasmic side of the FimD β-barrel—are also resolved in the 3D maps.

Conformer 2 is similar to the previously published crystal structure of the FimD–tip complex, in which FimF, FimG and FimH have polymerized into a tip fibre (that is, the N-terminal extensions of FimF and FimG are engaged in donor-strand exchange with FimG and FimH, respectively) and FimC–FimF has already transferred to the CTDs of FimD[20]. In conformer 2, the NTD of FimD makes no interactions with FimC–FimF and instead resides about 10 Å away in a pose in which it is ready to recruit the next incoming chaperone–subunit complex (Fig. 2b). In this conformation, the N-terminal tail of FimD is not visible, probably owing to flexibility, which is consistent with previously published structures of the disengaged NTD[13,14,20].

Conformer 1 reveals a previously undetected interaction between the NTD of the usher and CTD2. All of the periplasmic domains of the usher exhibit conformational changes compared to conformer 2 (Extended Data Fig. 4, Supplementary Video 2). The NTD of FimD shifts laterally by about 30 Å and rotates by about 45°, to interact with CTD2. This NTD shift leads to a loss of contact with the plug (Extended Data Fig. 5). The movement of the NTD causes the bound FimC–FimF

[1]Structural Biology Program, Van Andel Research Institute, Grand Rapids, MI, USA. [2]Department of Molecular Genetics and Microbiology, Stony Brook University, Stony Brook, NY, USA. [3]Center for Infectious Diseases, Stony Brook University, Stony Brook, NY, USA. [4]Department of Biochemistry and Cell Biology, Stony Brook University, Stony Brook, NY, USA. [5]David Van Andel Advanced Cryo-Electron Microscopy Suite, Van Andel Research Institute, Grand Rapids, MI, USA. [6]Present address: Department of Urology, Glickman Urological and Kidney Institute, Cleveland Clinic, Cleveland, OH, USA. *e-mail: David.Thanassi@stonybrook.edu; Huilin.Li@vai.org

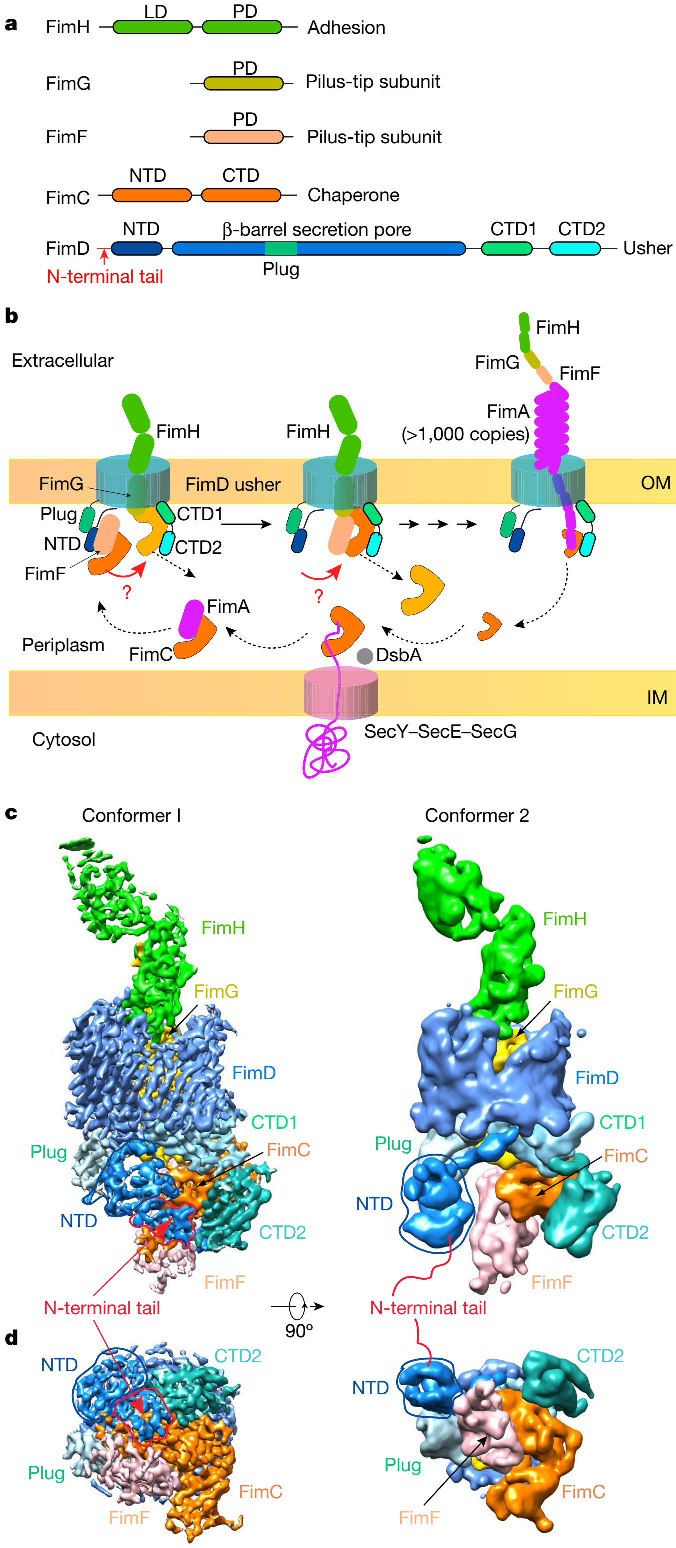


Fig. 1 | The missing link in pilus biogenesis, and cryo-EM of FimD–tip complex. a, FimD–tip components. LD, lectin domain; PD, pilin domain; N-terminal tail, the 24 residues preceding the folded NTD of FimD. **b**, A sketch of the pilus biogenesis pathway. The red arrows and question marks highlight a key unknown step—the handover of the chaperone–subunit from the NTD to the CTDs of the usher. DsbA catalyses disulfide bond formation in a nascent subunit, which is a prerequisite for subunit recognition by FimC[23]. OM, outer membrane; IM, inner membrane. **c**, **d**, The cryo-EM 3D map of conformer 1 at 4.0-Å resolution (**c**) and of conformer 2 at 5.1-Å resolution (**d**), respectively. Subunits are individually coloured as in **a**.

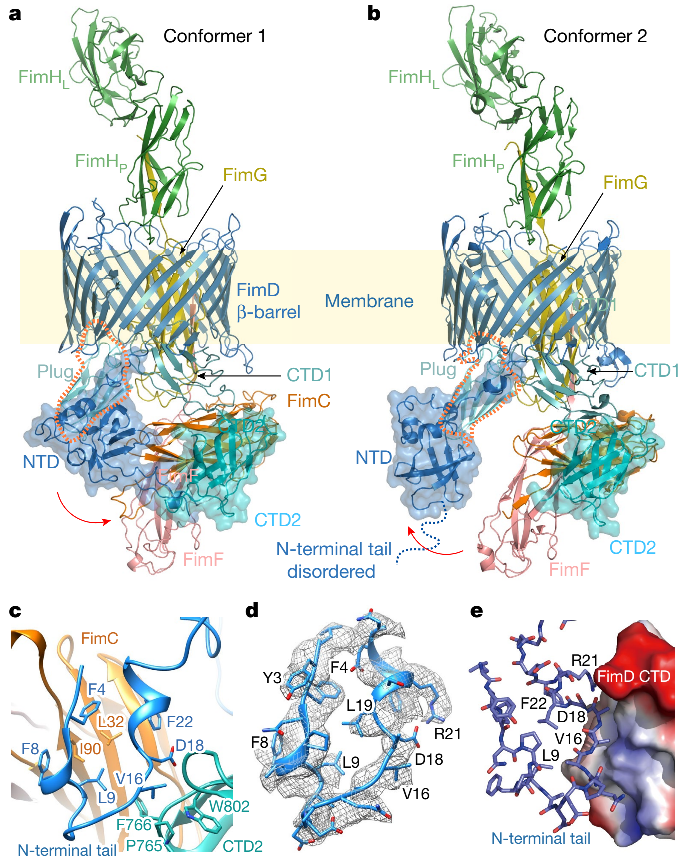


Fig. 2 | Atomic models of conformers 1 and 2. a, **b**, Cryo-EM structures of the FimD–tip complex in conformer 1 (**a**) and conformer 2 (**b**), coloured as in Fig. 1a. The dashed orange shape highlights the plug, which contacts the NTD in conformer 2 but loses contact in conformer 1 (see Extended Data Fig. 5). FimH$_L$, lectin domain of FimH; FimH$_P$, pilin domain of FimH. **c**, Interactions between N-terminal tail and CTD2 of FimD in cartoon view. **d**, Electron density for the N-terminal tail of FimD, with a local resolution of 4.0 Å. **e**, N-terminal tail of FimD in cartoon view, interacting with CTD2 of FimD in surface-charge view, ranging from positive (blue) to negative (red) charges.

to move upwards by about 5 Å. To accommodate the NTD and FimC–FimF movement, the remaining periplasmic components (the plug, CTD1 and CTD2) also undergo considerable rigid-body movements of approximately 10 Å. Owing to these movements, in conformer 1 the interactions between FimC–FimF and the CTDs of FimD are much weaker than those in conformer 2, consistent with the transitional nature of conformer 1 (Extended Data Fig. 5). Thus, conformer 1 captures the usher in the midst of the handover process, at a stage that appears to be immediately after incorporation of FimF into the nascent pilus tip; that is, FimF has already engaged in donor-strand exchange with FimG (Fig. 2a, b). We therefore conclude that subunit polymerization via the formation of donor-strand exchange precedes release of the incoming subunit from the NTD of the usher.

In conformer 1, the N-terminal tail of FimD folds into a U-shape to interact with CTD2 and FimC (Fig. 2c, d). The N-terminal tail binds to FimC via extensive hydrophobic interactions: Phe4, Phe8, Leu9 and Phe22 of the N-terminal tail interact with Leu32 and Ile90 of FimC. Previous crystal structures of the NTD of FimD in isolation showed that these same residues participate in the interface between the NTD and the FimC subunit[14,22]. Point mutations of these residues or deletions within the N-terminal tail disrupt pilus biogenesis[14,15] (Extended Data Table 2). Interactions between the N-terminal tail and CTD2 are of a mixed nature: Leu9 and Val16 are in proximity to, and may interact hydrophobically with, Pro765 and Phe766 of CTD2, and Asp18 may interact with Trp802 of CTD2. A FimD(L9E) mutant was unable to

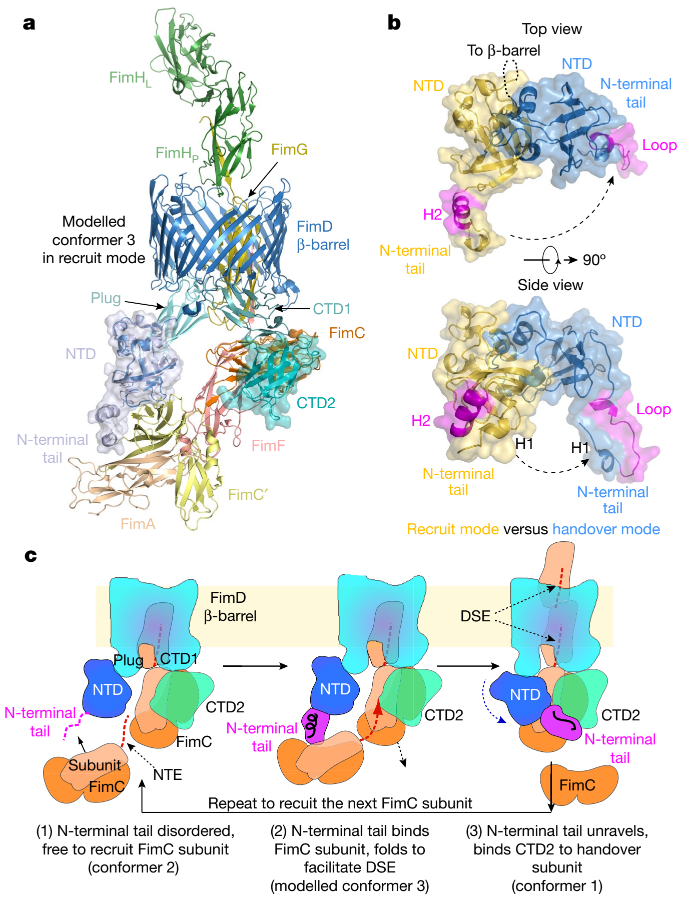


Fig. 3 | The N-terminal tail of FimD adopts three conformations during a subunit-incorporation cycle. a, Modelled conformer 3 in which FimA (wheat) bound to a FimC chaperone (yellow) is being recruited by the FimD–tip complex. In the recruitment phase, the N-terminal tail of FimD is folded as a helical motif. **b**, Comparison of the NTD and N-terminal tail of FimD in conformers 1 and 3 by superimposing the FimD β-barrel. H1, helix 1; H2, helix 2. **c**, A three-step handover mechanism of FimC subunit from the NTD to the CTDs of the usher, highlighting the three conformations of the N-terminal tail of FimD: disordered in step 1, helical in step 2 and a loop in step 3. DSE, donor-strand exchange; NTE, N-terminal extension.

assemble type 1 pili on the bacterial surface and FimD(P765E) and FimD(F766E) mutants exhibited partial pilus assembly defects, as assessed using a haemagglutination assay (Extended Data Table 2). Although Trp802 and Pro765 of CTD2 may not directly contact the N-terminal tail of FimD,these residues seem to form a hydrophobic network with Val16 of the N-terminal tail. The residues involved in contact between the NTD and CTD2 of FimD are well-conserved among usher proteins[5] (Extended Data Fig. 6).

FimF is the last subunit of the tip to be incorporated and is followed by the incorporation of many copies of FimA, which form the pilus rod. Using crystal structures of FimC–FimA and FimD$_{NTD}$–FimC–FimF[22,23], we built an atomic model of FimD–FimC–FimF–FimG–FimH–FimC–FimA (termed conformer 3) to show how FimA is recruited to the FimD–tip complex[20] (Fig. 3a). In this recruitment mode, the N-terminal tail of the NTD of FimD is folded into a two-helix motif[14,22]. Comparing conformers 1 and 3 reveals the nature of the conformational changes in FimD during the handover process (Fig. 3b): the folded core of the NTD undergoes a rigid-body movement and dissociates from FimC–FimF, and the FimD N-terminal tail maintains contact with FimC–FimF but changes from the helical motif to a loop, exposing several residues that can then interact with CTD2. Despite these changes, contacts between the N-terminal tail and FimC–FimF are similar in conformers 1 and 3. This is possible because the N-terminal tail is connected to the NTD by a flexible linker, allowing the NTD to rotate away from FimC–FimF while the N-terminal tail remains bound (Extended Data Fig. 7). Thus, the NTD is able to maintain contact with FimC–FimF throughout the handover process while allowing the CTDs access to the common binding surface on FimC; this permits the transfer of FimC–FimF to the CTDs.

Our work reveals a folding–unfolding cycle of the N-terminal tail of FimD: the N-terminal tail is disordered in conformer 2 (Fig. 2b), adopts a small helical motif in conformer 3 (Fig. 3a) and rearranges to an ordered loop in conformer 1 (Fig. 2a). Based on this observation, we propose a more-detailed model of the pilus assembly mechanism (Fig. 3c). In the recruitment mode, the N-terminal tail of FimD probably swings freely in search of an incoming chaperone–subunit complex. Once captured, the N-terminal tail folds into a helical motif to position the captured chaperone–subunit for donor-strand exchange with the previously recruited subunit on the CTDs, which leads to polymerization of the new subunit and displacement of the chaperone from the preceding subunit. The handover of the newly incorporated chaperone–subunit from the NTD to the CTDs is then driven by the higher affinity of the CTDs for the shared binding site on FimC[24], together with the formation of the new N-terminal-tail interface with CTD2. The formation of this interface may also enable CTD2 to facilitate the handover process, by destabilizing chaperone–subunit binding to the NTD[25]. The maintenance of N-terminal-tail binding to the newly recruited chaperone–subunit throughout the handover process allows transfer of the complex to the CTDs without complete release from the NTD. This provides a mechanism to impose directionality on subunit polymerization at the usher, ensuring outward translocation of the pilus fibre. Once the end of the growing pilus is handed over to the CTDs, the release of the N-terminal tail from CTD2 may be facilitated by the binding of the NTD to the plug domain, which resets the usher for a new cycle of chaperone–subunit recruitment and polymerization. Thus, the catalytic cycle involves a meeting between the two extremities of the usher.

The capture of conformers 1 and 2 in the same solution indicates that they exist in a dynamic equilibrium. It is conceivable that conformer 1, in which the NTD swings over to bind CTD2, stabilizes the FimC chaperone against dissociation from the end of the growing pilus fibre. Dissociation of the chaperone would risk release of the pilus fibre and diffusion through the usher pore into the extracellular medium. This stabilization function of the NTD of the usher would be absent in conformer 2. However, because the two conformers are in equilibrium, conformer 2 probably reverts back to conformer 1 faster than FimC spontaneously dissociates from the pilus end. In this regard, conformer 1 may have a dual function: providing a mechanism for handing over the end of growing pilus from the NTD to the CTDs of the usher, and stabilizing the growing pilus fibre against diffusion away from the usher.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0587-z.

Received: 5 February 2018; Accepted: 15 August 2018;
Published online 3 October 2018.

1. Flores-Mireles, A. L., Walker, J. N., Caparon, M. & Hultgren, S. J. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nat. Rev. Microbiol.* **13**, 269–284 (2015).
2. Thanassi, D. G., Saulino, E. T. & Hultgren, S. J. The chaperone/usher pathway: a major terminal branch of the general secretory pathway. *Curr. Opin. Microbiol.* **1**, 223–231 (1998).
3. Geibel, S. & Waksman, G. The molecular dissection of the chaperone–usher pathway. *Biochim. Biophys. Acta* **1843**, 1559–1567 (2014).
4. Zav'yalov, V., Zavialov, A., Zav'yalova, G. & Korpela, T. Adhesive organelles of Gram-negative pathogens assembled with the classical chaperone/usher machinery: structure and function from a clinical standpoint. *FEMS Microbiol. Rev.* **34**, 317–378 (2010).
5. Nuccio, S. P. & Bäumler, A. J. Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek. *Microbiol. Mol. Biol. Rev.* **71**, 551–575 (2007).

6. Mulvey, M. A. et al. Induction and evasion of host defenses by type 1-piliated uropathogenic *Escherichia coli*. *Science* **282**, 1494–1497 (1998).
7. Sauer, F. G. et al. Chaperone-assisted pilus assembly and bacterial attachment. *Curr. Opin. Struct. Biol*. **10**, 548–556 (2000).
8. Choudhury, D. et al. X-ray structure of the FimC–FimH chaperone–adhesin complex from uropathogenic *Escherichia coli*. *Science* **285**, 1061–1066 (1999).
9. Sauer, F. G. et al. Structural basis of chaperone function and pilus biogenesis. *Science* **285**, 1058–1061 (1999).
10. Zavialov, A. V. et al. Structure and biogenesis of the capsular F1 antigen from *Yersinia pestis*: preserved folding energy drives fiber formation. *Cell* **113**, 587–596 (2003).
11. Sauer, F. G., Pinkner, J. S., Waksman, G. & Hultgren, S. J. Chaperone priming of pilus subunits facilitates a topological transition that drives fiber formation. *Cell* **111**, 543–551 (2002).
12. Nishiyama, M., Ishikawa, T., Rechsteiner, H. & Glockshuber, R. Reconstitution of pilus assembly reveals a bacterial outer membrane catalyst. *Science* **320**, 376–379 (2008).
13. Phan, G. et al. Crystal structure of the FimD usher bound to its cognate FimC–FimH substrate. *Nature* **474**, 49–53 (2011).
14. Nishiyama, M. et al. Structural basis of chaperone–subunit complex recognition by the type 1 pilus assembly platform FimD. *EMBO J*. **24**, 2075–2086 (2005).
15. Ng, T. W., Akman, L., Osisami, M. & Thanassi, D. G. The usher N terminus is the initial targeting site for chaperone–subunit complexes and participates in subsequent pilus biogenesis events. *J. Bacteriol*. **186**, 5321–5331 (2004).
16. Saulino, E. T., Thanassi, D. G., Pinkner, J. S. & Hultgren, S. J. Ramifications of kinetic partitioning on usher-mediated pilus biogenesis. *EMBO J*. **17**, 2177–2185 (1998).
17. Volkan, E. et al. Domain activities of PapC usher reveal the mechanism of action of an *Escherichia coli* molecular machine. *Proc. Natl Acad. Sci. USA* **109**, 9563–9568 (2012).
18. Nishiyama, M. & Glockshuber, R. The outer membrane usher guarantees the formation of functional pili by selectively catalyzing donor-strand exchange between subunits that are adjacent in the mature pilus. *J. Mol. Biol*. **396**, 1–8 (2010).
19. Rose, R. J. et al. Unraveling the molecular basis of subunit specificity in P pilus assembly by mass spectrometry. *Proc. Natl Acad. Sci. USA* **105**, 12873–12878 (2008).
20. Geibel, S., Procko, E., Hultgren, S. J., Baker, D. & Waksman, G. Structural and energetic basis of folded-protein transport by the FimD usher. *Nature* **496**, 243–246 (2013).
21. Allen, W. J., Phan, G., Hultgren, S. J. & Waksman, G. Dissection of pilus tip assembly by the FimD usher monomer. *J. Mol. Biol*. **425**, 958–967 (2013).
22. Eidam, O., Dworkowski, F. S., Glockshuber, R., Grütter, M. G. & Capitani, G. Crystal structure of the ternary FimC–FimF$_t$–FimD$_N$ complex indicates conserved pilus chaperone–subunit complex recognition by the usher FimD. *FEBS Lett*. **582**, 651–655 (2008).
23. Crespo, M. D. et al. Quality control of disulfide bond formation in pilus subunits by the chaperone FimC. *Nat. Chem. Biol*. **8**, 707–713 (2012).
24. Werneburg, G. T. et al. The pilus usher controls protein interactions via domain masking and is functional as an oligomer. *Nat. Struct. Mol. Biol*. **22**, 540–546 (2015).
25. Volkan, E. et al. Molecular basis of usher pore gating in *Escherichia coli* pilus biogenesis. *Proc. Natl Acad. Sci. USA* **110**, 20741–20746 (2013).

Acknowledgements Cryo-EM data were collected at the David Van Andel Advanced Cryo-Electron Microscopy Suite at the Van Andel Research Institute. We thank X. Meng for help with data collection. This study was supported by the US National Institutes of Health R01 grants GM062987 (to D.G.T. and H.L.) and GM111742 (to H.L.) and Van Andel Research Institute (to H.L.).

Reviewer information *Nature* thanks A. Dessen, J. Rubinstein and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.G.T. and H.L. conceived and designed experiments. M.D., N.H., S.S. and G.T.W. carried out biochemical and molecular biology experiments. M.D., Z.Y. and G.Z. performed cryo-EM experiments. M.D., Z.Y. and H.Y. performed image processing and atomic modelling. M.D., Z.Y., H.Y, D.G.T. and H.L. analysed the data. M.D., Z.Y., H.Y., D.G.T. and H.L. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0587-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0587-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.G.T. or H.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

FimD–tip complex expression and purification. The His-tagged FimD–tip complex (FimD–FimC–FimF–FimG–FimH; FimDCFGH) was expressed in *E. coli* Tuner competent cells as previously described[26]. In brief, 60 l of bacteria was grown at 37 °C with aeration in LB supplemented with 50 μg/ml kanamycin and 25 μg/ml chloramphenicol. At $OD_{600} = 0.6$, FimDCFGH was induced with 25 μM IPTG and 0.05% (w/v) arabinose overnight at 20 °C. The bacteria were collected and resuspended using a homogenizer into 25 mM Tris-HCl (pH 8.0), 100 mM NaCl, 1 mM β-ME. The cells were disrupted with sonication or by a microfluidizer. The volume was made to 1.8 l using 25 mM Tris-HCl (pH 8.0), 100 mM NaCl, 1 mM β-ME. The disrupted cells were spun (10,000*g*, 20 min, 4 °C) to separate cellular debris and unbroken cells. Subsequently, 0.5% (w/v) sarkosyl was added to the supernatant to solubilize the inner membrane (stirring, 20 min, room temperature). The outer membrane was isolated by ultracentrifugation (100,000*g*, 50 min, 4 °C). The outer membrane was washed three times using 25 mM Tris-HCl (pH 8.0), 100 mM NaCl, 1 mM β-ME and ultracentrifugation was used to isolate the pure outer membrane complex. The outer membrane was resuspended using 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 15% glycerol, 10 mM $MgCl_2$ and protease inhibitors. FimDCFGH was extracted by adding 30 mg of lysozyme and 1% (w/v) *n*-dodecyl β-D-maltoside (DDM; Anatrace) (stirring, overnight, 4 °C). The insoluble material was separated out by ultracentrifugation (100,000*g*, 1 h, 4 °C). The solubilized FimDCFGH was applied to a Ni-NTA column and washed several times with 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 0.5% DDM. The Ni-NTA column was washed with 50 mM buffer containing imidazole for removal of impurities. FimDCFGH was eluted using buffer containing 250 mM imidazole. Purified FimDCFGH was concentrated using a Centricon 100 concentrator. The concentrated sample was applied on a size-exclusion chromatography column Superdex 200 (GE Healthcare), equilibrated with 25 mM Tris-HCl (pH 8.0), 200 mM NaCl, 0.5% DDM. The fractions of the size-exclusion chromatography were analysed on SDS–PAGE (Extended Data Fig. 2a, b). The non-aggregated fractions containing the FimDCFGH complex were concentrated.

Cryo-EM. To prepare cryo-EM grids, a 3-μl FimD–tip sample was applied to glow-discharged C-flat 1.2/1.3 holey carbon grids, incubated for 10 s at 6 °C and 95% humidity, blotted for 3 s and then plunged into liquid ethane using an FEI Vitrobot IV. In C-flat R1.2/1.3 holey carbon film grids, the FimD–tip particles distributed well with no aggregation problems. The grids were loaded into an FEI Titan Krios electron microscope operated at a high tension of 300 KV and images were collected semi-automatically with EPU under low-dose mode at a nominal magnification of 130,000× and a pixel size of 1.09 Å per pixel. A Gatan K2 summit direct electron detector was used in super-resolution mode for image recording with an under-focus range from 1.5 to 2.5 μm. A Bioquantum energy filter installed in front of the K2 detector was operated in the zero-energy-loss mode with an energy-slit width of 20 eV. The dose rate was 10 electrons per $Å^2$ per second and total exposure time was 6 s. The total dose was divided into a 30-frame movie, such that each frame was exposed for 0.2 s.

Image processing and 3D reconstruction. Approximately 12,000 raw movie micrographs were collected. The movie frames were first aligned and superimposed by the program Motioncorr 2.0[27]. Contrast transfer function parameters of each aligned micrograph were calculated using the program CTFFIND4[28]. All of the remaining steps, including particle auto-selection, 2D classification, 3D classification, 3D refinement and density map post-processing were performed using Relion-2.1[29]. The template for automatic picking was generated from a 2D average of about ~10,000 manually picked particles in different views. Automatic particle selection was performed for the entire dataset, and 1,534,108 particles were initially picked. Selected particles were carefully inspected; 'bad' particles were removed, some initially missed 'good' particles were re-picked and the remaining good particles were sorted by similarity to the 2D references, in which the bottom 10% of particles with the lowest *z*-scores were removed from the particle pool. Two-dimensional classification of all good particles was performed and particles in the classes with unrecognizable features by visual inspection were removed. A total of 758,698 particles was used for further 3D classification. Five 3D models were derived from the dataset, and the two best models were chosen for final refinement (Extended Data Fig. 2c–e). The other three models were distorted and those particles were discarded. The final two datasets have 250,370 and 166,913 particles, respectively. They were used for further 3D refinement, resulting in the 4.0-Å and 5.1-Å 3D density map. The resolution of the map was estimated by the gold-standard Fourier shell correlation, at the correlation cut-off value of 0.143. The 3D density map was sharpened by applying a negative B-factor of −230 and −241 $Å^2$, respectively (Extended Data Fig. 3a, b).

Atomic modelling, refinement and validation. The modelling of two conformations was based on the crystal structure of FimD–FimC–FimF–FimG–FimH (RCSB Protein Data Bank (PDB) ID 4J3O). For conformation 1, the complex structure (PDB ID 4J3O) was split into individual subunits and individually docked into the electron microscopy map using Chimera[30]. The N-terminal tail region (2–27 amino acids) of FimD was absent from the FimDCFGH structure and its model was obtained from structure of FimD NTD–FimC–FimF structure (PDB ID 3BWU), which was fitted into the map using Chimera. The initial modelling was followed by further manual adjustments using COOT[31], guided by residues with bulky side chains such as Arg, Phe, Tyr and Trp (Extended Data Fig. 3c). The improved model was refined in real space against electron microscopy densities using the phenix.real_space_refine module in PHENIX[32]. For conformation 2, Chimera was used to rigid-body-dock the whole structure of FimD–tip (PDB ID 4J3O) into the corresponding electron microscopy map, fitting well into the densities. Only FimH was separated and individually rigid-body fitted into the density using Chimera. Owing to the low resolution of this conformation, the structure was not subject to further refinement. The FimD–tip conformer 3 was modelled by (1) superimposing the NTD in the FimD NTD–FimC–FimF structure (PDB ID: 3BWU) with the NTD of FimD in FimD–tip conformer 2, and (2) superimposing FimC in the FimC–FimA structure (PDB ID: 4DWH) with FimC in the FimD NTD–FimC–FimF complex. Finally, the atomic model of conformer 1 was validated using MolProbity[33]. Structural figures were prepared in Chimera and Pymol (https://www.pymol.org)

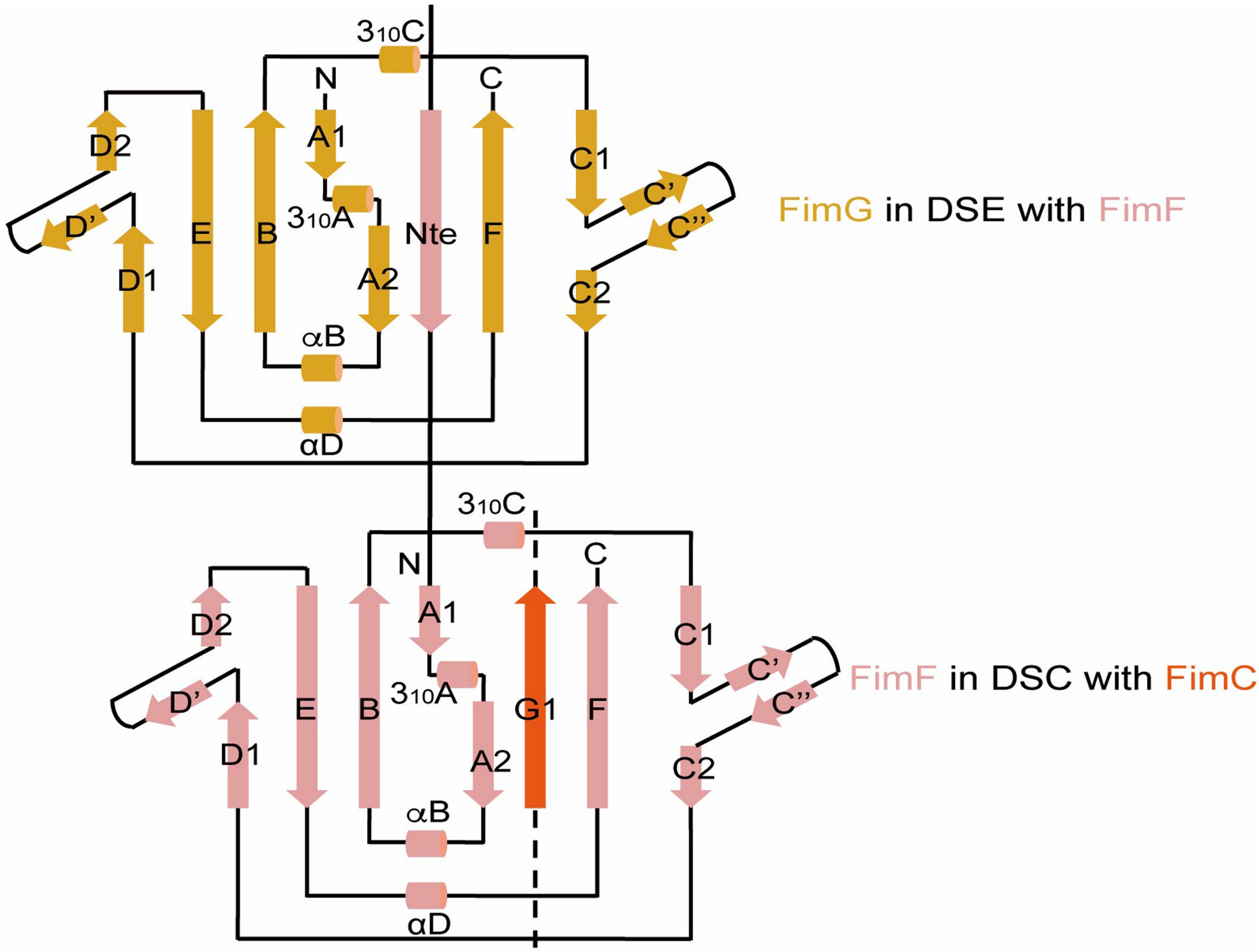
FimD mutagenesis and haemagglutination assay. For construction of the FimD(L9E), FimD(V16E), FimD(P765E) and FimD(F766E) mutants, plasmid pNH213[24] was mutated using the QuikChange Site-Directed Mutagenesis Kit (Stratagene) and primers as follows: L9E, 5′-CCGACCTCTATTTAATCCGC GCTTTGAAGCGGATGATCC-3′; V16E, 5′-GGATGATCCCCAGGCTGAGG CCGATTTATCG-3′; P765E, 5′-CCACAATAATAAGCCGCTGGAGTTTGGG GCGATGGTGAC-3′; and F766E, 5′-CCACAATAATAAGCCGCTGCCGGAGG GGGCGATGGTGAC-3′. Proper expression and folding of the FimD mutants in the bacterial outer membrane were determined by heat-modifiable mobility, as described previously[34]. Comparison of the mutants with wild-type FimD for ability to assemble type 1 pili on the bacterial surface was performed using a haemagglutination assay, as described previously[34]. Haemagglutination titres were recorded visually as the greatest fold dilution of bacteria able to agglutinate guinea pig red blood cells (Colorado Serum). Titres were calculated from three independent experiments of three replicates each.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

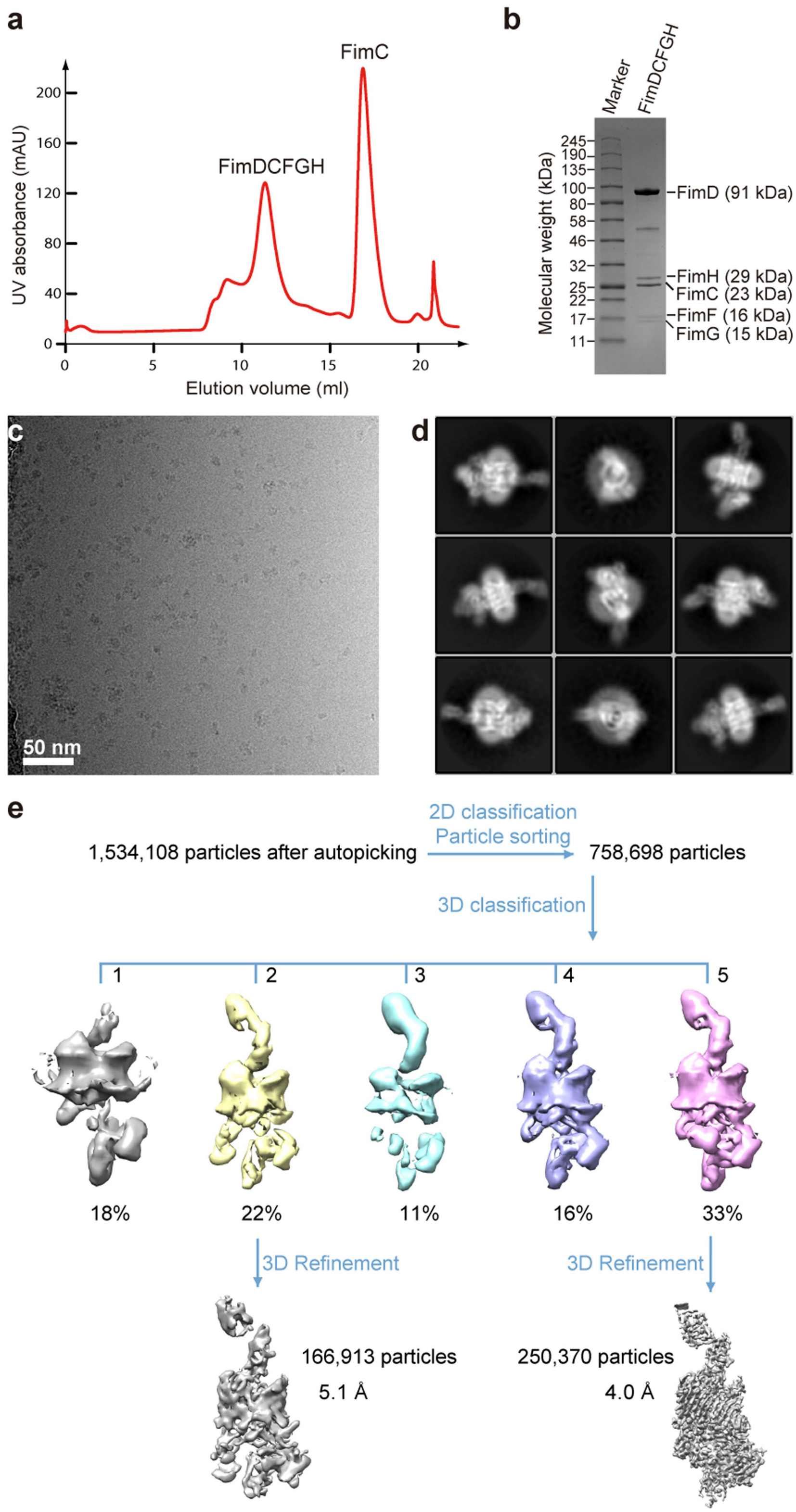
Data availability

The cryo-EM 3D maps of the FimD–tip (FimDCFGH) complex have been deposited at the Electron Microscopy Data Bank database with accession codes EMD-8953 (conformer 1) and EMD-8954 (conformer 2), and their corresponding atomic models were deposited at the PDB with accession codes 6E14 (conformer 1) and 6E15 (conformer 2), respectively.

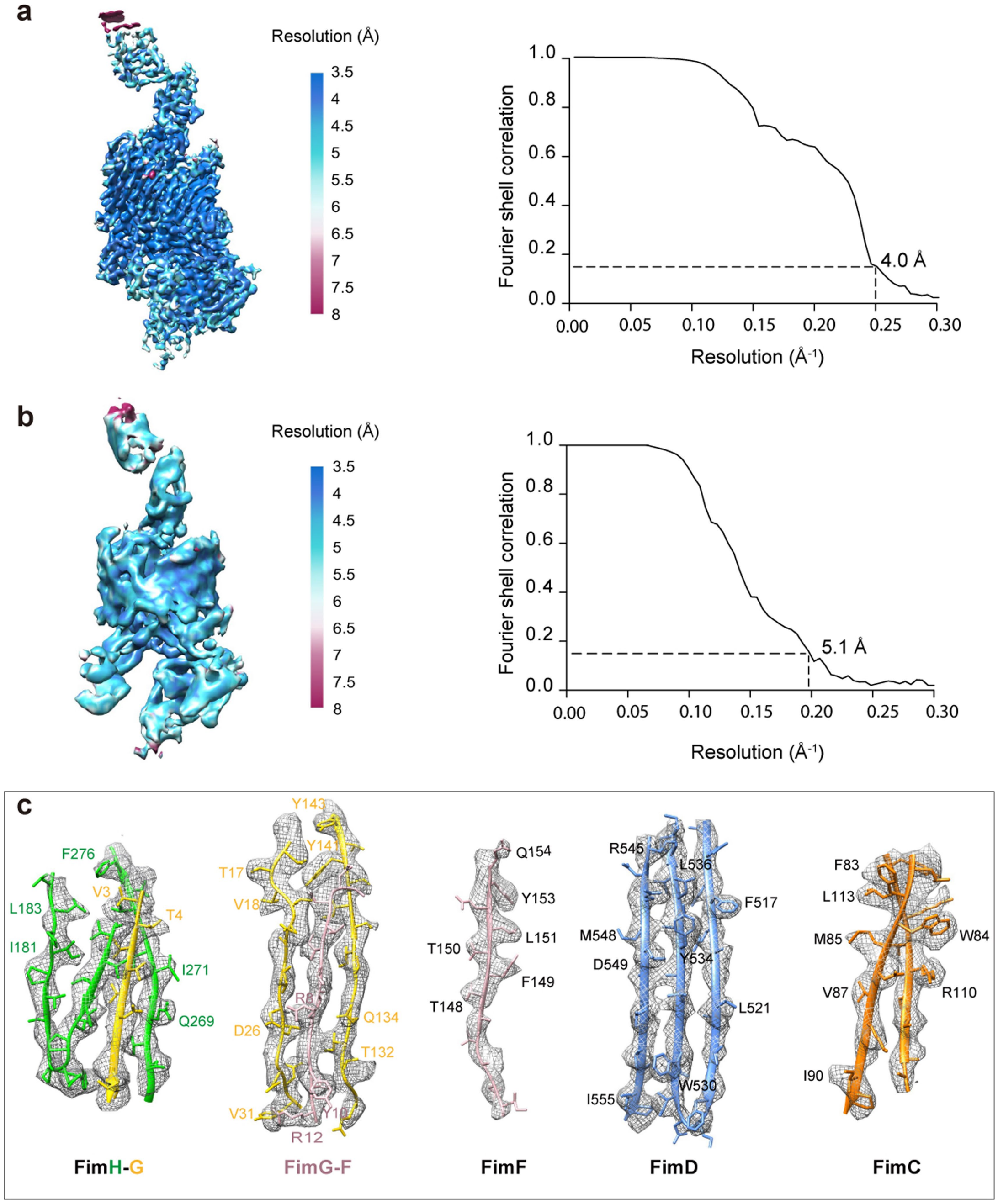
26. Remaut, H. et al. Fiber formation across the bacterial outer membrane by the chaperone/usher pathway. *Cell* **133**, 640–652 (2008).
27. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
28. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol*. **192**, 216–221 (2015).
29. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol*. **180**, 519–530 (2012).
30. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem*. **25**, 1605–1612 (2004).
31. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
32. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
33. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
34. Henderson, N. S., Ng, T. W., Talukder, I. & Thanassi, D. G. Function of the usher N-terminus in catalysing pilus assembly. *Mol. Microbiol*. **79**, 954–967 (2011).



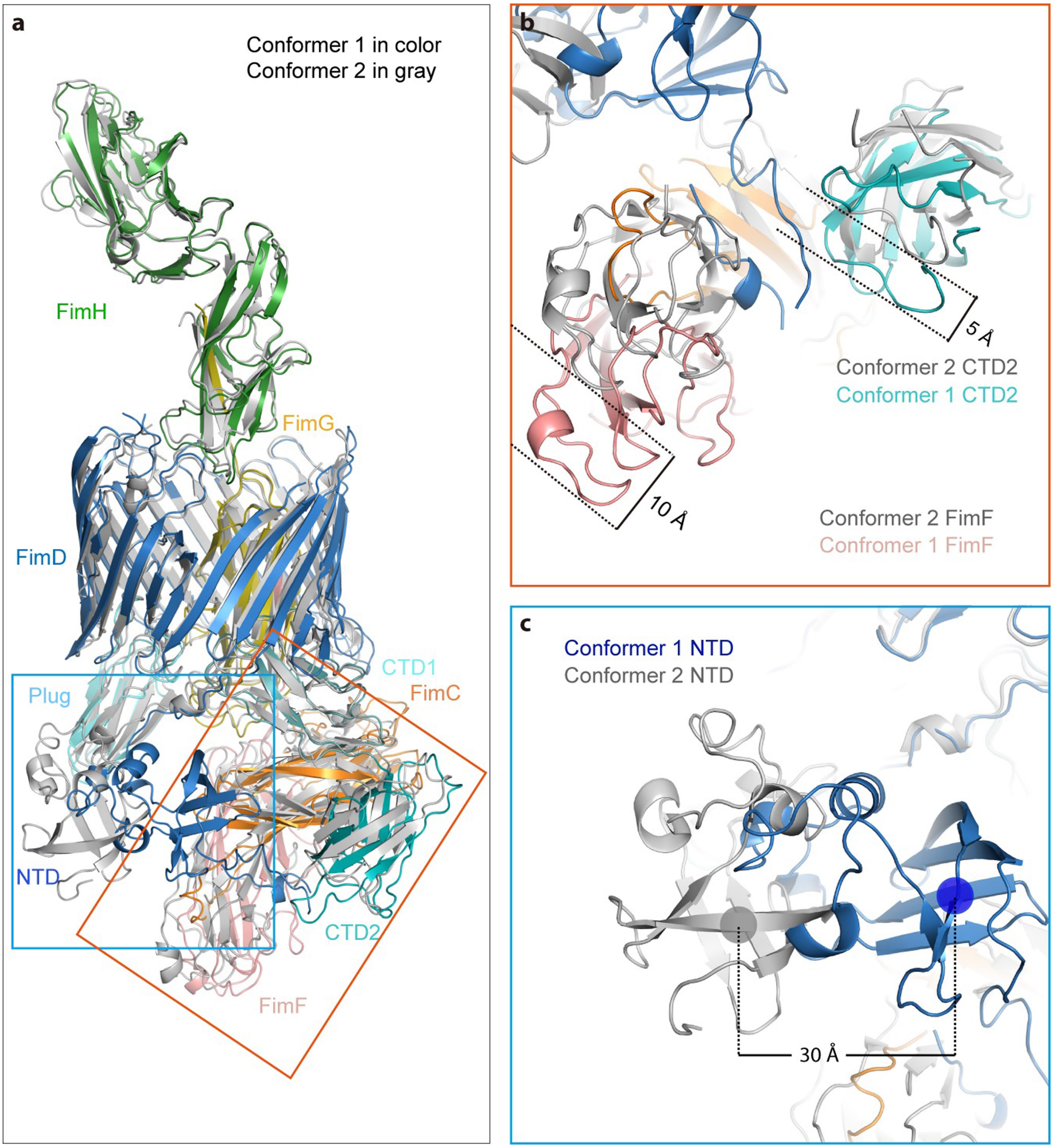
Extended Data Fig. 1 | Pilus assembly occurs via donor-strand complementation and donor-strand exchange. This sketch is based on the crystal structure of FimD–FimC–FimF–FimG–FimH (PDB ID 4J3O). Donor-strand complementation (DSC): pilus subunits (FimF, in this case) have an immunoglobulin-like structure in which the C-terminal G strand is missing. In the periplasm, the chaperone FimC donates its G1 strand to complete the subunit fold, but in a non-canonical parallel orientation with the subunit F strand. Donor-strand exchange (DSE): the donor strand of the FimC chaperone in the previous subunit (FimG, in this case) is replaced by the N-terminal extension of the incoming subunit (FimF), completing the FimG-subunit immunoglobulin fold in a canonical, anti-parallel orientation.



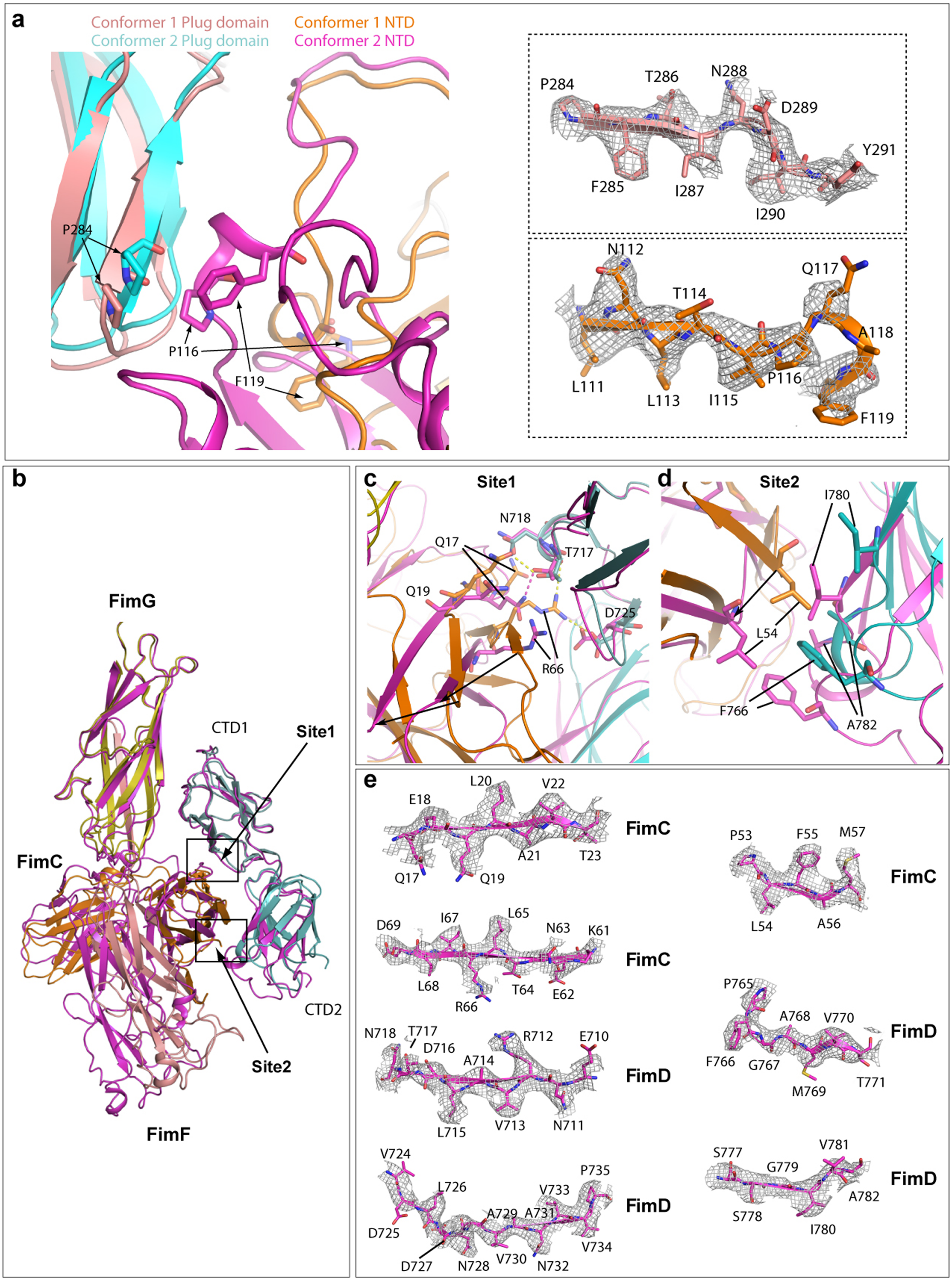
Extended Data Fig. 2 | Cryo-EM of FimD–tip complex. a, The gel filtration profile of FimD–tip complex from a Superdex 200 10/300GL column. **b**, Coomassie blue SDS–PAGE gel of the peak fraction, which shows the presence of all subunits of the purified FimD–tip complex. Similar sample preparations by gel filtration and SDS–PAGE examination were carried out more than three times. **c**, A raw cryo-EM micrograph of the purified FimD–tip complexes embedded in vitreous ice. A total of 12,000 such micrographs was recorded. **d**, Selected 2D class averages showing the presence of many different views and well-resolved structural features. Over 750,000 raw particles contributed to final 2D class averages. **e**, Three-dimensional classification scheme. Over 1 million raw particles were selected from drift-corrected electron micrographs. Two- and three-dimensional classification resulted in two 3D maps that were of the expected shape, and the structure appeared complete; the other three maps were either partial structures or distorted. Refinement with about 250,370 particles led to the 4.0-Å resolution 3D map of conformer 1, and refinement with about 166,913 particles led to the 5.1-Å resolution 3D map of conformer 2.



Extended Data Fig. 3 | Resolution estimation and selected regions of the 3D electron microscopy maps. a, Local-resolution estimation and the Gold-standard Fourier shell correlation estimation at the 0.143 correlation threshold of the FimD–tip in conformer 1. **b**, As in **a**, but for conformer 2. **c**, Model fitting in the FimD–tip density map. Selected densities for FimH–FimG, FimG–FimF, FimF, FimD and FimC of conformer 1 are shown. Amino acids with clear side-chain densities are indicated.

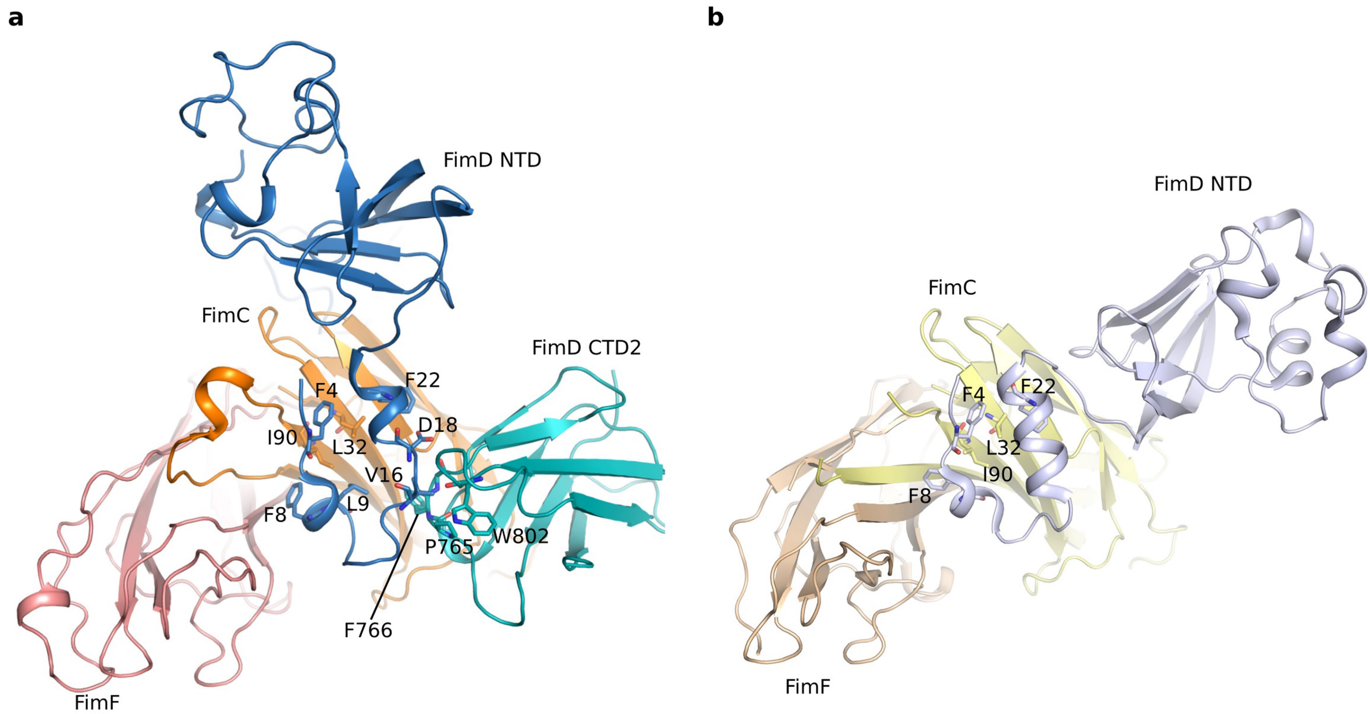


Extended Data Fig. 4 | Comparison between conformer 1 and conformer 2 of the FimD–tip complex. a, Overlap of conformer 1 (in colour cartoon) with that of conformer 2 (in grey cartoon). The movement of the NTD is labelled in the blue box, and the movement of CTD2 and FimF is labelled in the orange box. **b**, Comparison of FimF and CTD2 of conformer 1 with those of conformer 2. FimF and CTD2 shift 10 Å and 5 Å, respectively. **c**, Comparison of the NTD of conformer 1 with that of conformer 2. The NTD shifts laterally by about 30 Å and rotated by about 45°.



Extended Data Fig. 5 | Comparison of interaction between plug domain and NTD of FimD and between the CTDs of FimD and FimC–FimF in conformers 1 and 2 by superimposing the two conformations. **a**, Left, the plug domain and NTD of conformer 1 are coloured in salmon and orange, respectively; plug domain and NTD of conformer 2 are coloured in cyan and magenta, respectively. Right, electron densities of regions involved in the interaction in conformer I are shown for the FimD plug (top) and FimD NTD (bottom). Some amino acids have clear side-chain densities. **b**, Superimposition of FimD CTDs–FimC–FimF in conformer 1 (magenta) and conformer 2 (coloured as in Fig. 3). **c**, Detailed interactions in site 1 (marked in **b**). Extensive interactions are present in conformer 2. Much weaker interactions are present in conformer 1 (between Q17 and T717). **d**, Detailed interactions in site 2 (marked in **b**). In conformer 2, hydrophobic interactions exist between FimC L54 and FimD F766, and between I780 and A782 of FimD. Much weaker interactions are present in conformer 1. **e**, The electron densities in regions involved in interactions (in sites 1 and 2) between FimC and FimD in conformer 1. Some amino acids have side-chain densities.

Extended Data Fig. 6 | Sequence conservation of the interacting interface at the two extreme termini of the FimD usher. Residues involved in the interaction between the NTD of FimD and FimC are labelled with circles, and residues involved in the interaction between the NTD and CTD2 of FimD are labelled with triangles.



Extended Data Fig. 7 | The interactions between the NTD of FimD and FimC–FimF. a, **b**, The interactions between the NTD of FimD and FimC–FimF are shown for conformer 1 (**a**) and modelled conformer 3 (**b**). There are no interactions between the NTD of FimD and the FimF subunit in either conformer. The interactions between the N-terminal tail of FimD and the FimC chaperone are essentially the same in the two conformers.

Extended Data Table 1 | Cryo-EM data collection and model statistics of FimD–tip complex

	Conformer 1 (EMD-8953) (PDB 6E14)	**Conformer 2** (EMD-8954) (PDB 6E15)
Data collection and processing		
Magnification	130,000	130,000
Voltage (kV)	300	300
Electron dose (e^-/Å^2)	50	50
Defocus range (μm)	-1.5 – -2.5	-1.5 – -2.5
Pixel size (Å)	1.09	1.09
Symmetry imposed	C1	C1
Initial particle images (no.)	758,698	758,698
Final particle images (no.)	250,370	166,913
Map resolution (Å)	4.0	5.1
FSC threshold	0.143	0.143
Map resolution range (Å)	3.5 – 5.0	4.0 – 6.0
Refinement		
Initial model used (PDB code)	4J3O	4J3O
Map sharpening B factor (Å^2)	-230	-241
Model composition		
Non-hydrogen atoms	11995	11732
Protein and DNA residues	1572	1536
Ligands	0	0
R.m.s. deviations		
Bond lengths (Å)	0.007	
Bond angles (°)	1.38	
Validation		
MolProbity score	1.52	
Clashscore	3.92	
Poor rotamers (%)	0.23	
Ramachandran plot		
Favored (%)	95.04	
Allowed (%)	4.71	
Disallowed (%)	0.25	

Extended Data Table 2 | Effects of mutations in N-terminal tail and CTD2 of FimD on pilus assembly

FimD	Agglutination	Reference
wild-type	128±0[a], +[b]	This study, [14]
F4A	0, –	[14,34]
F8A	+/-	[14]
L9E	0±0	This study
Δ1-11	–	[14]
V16E	128±0	This study
F22A	–	[14]
P765E	64±0	This study
F766E	64±0	This study

[a]Haemagglutination titres represent the highest-fold dilutions of bacteria that are able to agglutinate guinea pig red blood cells. The values report the effects of the usher mutations on overall bacterial pilus production. A titre of 128 equals a sevenfold dilution and a titre of 64 equals a sixfold dilution. Titres were calculated from three independent experiments of three replicates each; all values for each of the experiments and replicates were identical. The functional defects of the FimD(L9E), FimD(P765E) and FimD(F766E) mutants were not due to changes in expression or folding in the outer membrane, as determined by a heat-modifiable mobility assay.
[b]Ability of bacteria to agglutinate yeast cells: +, strong agglutination; ±, weak agglutination; −, no agglutination.

Re-evaluating the p7 viroporin structure

ARISING FROM B. OuYang et al. *Nature* **498**, 521–525 (2013); https://doi.org/10.1038/nature12283

The hepatitis C virus (HCV) p7 viroporin is a membrane protein required for virus propagation in vivo that assembles into hexamers and heptamers in membranes, exhibits ion channel activity, and is an attractive target against HCV infection[1]. OuYang and colleagues reported an oligomeric structure of p7 solubilized in dodecylphosphocholine (DPC) detergent, with unexpected features[2]. Here we show that p7 is monomeric in the conditions that were used to determine its oligomeric structure and that the data presented as evidence for intermolecular contacts is likely to arise from incomplete protein deuteration. We conclude that p7 is monomeric under NMR conditions, and that the oligomeric structure proposed by OuYang et al.[2] is artefactual. There is a Reply to this Comment by Chen, W. et al. *Nature* **562**, https://doi.org/10.1038/s41586-018-0562-8 (2018).

Unexpected features of the p7 oligomeric structure[2] include: (i) the His17 side chain, known to be involved in ion conduction[3,4], points outward towards the membrane bilayer; (ii) the orientation of the best fit of the oligomeric structure to the electron microscopy envelope contradicts antibody binding data[5]; and (iii) a short outer transmembrane helix exposes polar residues to the hydrophobic region of the membrane so that the structure cannot be accommodated in lipid bilayers without large structural rearrangements or membrane thinning[6,7] (Extended Data Fig. 1a, b).

We expressed, purified, and reconstituted into dodecylphosphocholine (DPC) ^{15}N-labelled protein corresponding to p7 of the genotype 5a isolate EUH1480, which contains five amino acid substitutions (p7(5a)[2], hereafter referred to as p7). Overlays of backbone ^{1}H–^{15}N-correlation NMR spectra confirm that the solution conditions and protein conformation are similar to those studied previously[2,8] (Extended Data Fig. 2a). High-quality backbone spectra of fully protonated p7 in DPC could be recorded at 37 °C using conventional, heteronuclear single quantum correlation (HSQC)-based experiments (Extended Data Fig. 2b), which is unexpected for a protein complex of 60–80 kDa[9] (the molecular mass of the protein oligomer and the associated detergent micelle). The spectra were unchanged between 30 °C and 37 °C, indicating that the protein adopted a similar conformation over this temperature range (Extended Data Fig. 2c).

Two methods were used to obtain information on the size of the p7–detergent-micelle complex. First, we estimated the effective protein mass using the rotational correlation time derived from ^{15}N R_1 and R_2 relaxation rates. The rotational correlation time, 10.1 ns at 37 °C, corresponds to 39.3 kDa, similar to a hexameric p7 (approximately 41 kDa) in the absence of a detergent micelle. Alternatively, the same rotational correlation time corresponds to monomeric p7 and a micelle of around 70 detergent molecules. We then applied size-exclusion chromatography coupled to multi-angle light scattering (SEC–MALS), which decomposes masses into protein and detergent contributions, and is reliable for proteins such as p7, which has a high molar extinction coefficient (17,990 M^{-1} cm^{-1}) and a relatively high light scattering intensity owing to the detergent micelle[10]. P7 samples reconstituted as described[2] and analysed using SEC–MALS in conditions similar to those used for NMR (50 mM DPC; Fig. 1a) or electron microscopy studies (3 mM DPC but a higher protein concentration; Fig. 1b) indicated protein masses consistent with monomers (6.4 ± 1.1 kDa and 8.4 ± 0.7 kDa, respectively) associated with detergent micelles of 23.5 ± 0.5 kDa and 42.4 ± 3.7 kDa, respectively (Extended Data Table 1). A third sample prepared using a different protocol[8] (50 mM DPC), indicated a protein mass of 7.8 ± 1.1 kDa associated with a detergent micelle of 24.7 ± 0.6 kDa (Extended Data Fig. 3a). Moreover, NMR spectra matching exactly the protein and detergent concentrations at the monomeric SEC–MALS peak summit (5 μM protein and 50 mM DPC) show good overlap with the spectrum published in ref. [2] (Extended Data Fig. 2d).

The SEC–MALS results are confirmed by comparison with p7 solubilized in sodium dodecyl sulfate (SDS) at 'low' (less than two times the critical micelle concentration (CMC)) and 'high' (at least ten times the CMC) detergent concentrations. SDS is strongly denaturing, and p7 runs as a monomer on SDS–PAGE[2]. Results for p7 in high (80 mM) and low (10 mM) SDS concentrations are consistent with a monomer, as expected, and similar to the results with DPC, in that low concentrations of SDS lead to an increase in detergent-micelle size with no change in oligomerization state (Extended Data Fig 3b–e and Extended Data Table 1). The same trend in micelle size with detergent concentration, albeit less marked, is observed in the absence of protein (Extended Data Fig. 3f–i and Extended Data Table 2).

OuYang et al.[2] identified putative intermolecular nuclear Overhauser effects (NOEs) from a 3D ^{15}N-edited NOE spectroscopy with transverse relaxation-optimized spectroscopy (NOESY–TROSY) experiment, without filtering for ^{13}C-attached protons, on a sample containing a 1:1 mixture of ^{15}N–^{2}H- and ^{13}C-labelled proteins. This experiment can provide intermolecular NOEs between exchangeable amide protons of the ^{15}N–^{2}H-labelled protein and non-exchangeable aliphatic protons of the ^{13}C-labelled protein. Artefactual cross-peaks can arise from incomplete deuteration of the ^{15}N–^{2}H-labelled protein, and a control NOESY experiment must be recorded on an unmixed but otherwise identical sample of ^{15}N–^{2}H-labelled protein, since commercial sources of D_2O and deuterated glucose contain at least small amounts of protons and additional protons can be carried over from starter cultures and humidity in the air. OuYang et al.[2] did not report control experiments, but comparison of the mixed-label NOESY spectra with a NOESY spectrum collected on fully protonated protein can indicate whether the NOEs are consistent with trace protonation of the ^{15}N–^{2}H-labelled protein. The NOESY spectra were aligned in the indirect ^{1}H dimensions such that the amide-proton chemical shifts, which exhibit relatively small deuterium-isotope shifts, were consistent. Seven NOEs were identified by OuYang et al.[2] as unambiguously intermolecular, and for each of these that could be identified in their mixed-label NOESY spectrum, there is a corresponding strong peak in their fully protonated sample NOESY close to the position expected for an intra-residue proton (Fig. 2a). The mixed-label NOEs were slightly shifted up-field on the order of 0.01 ppm, consistent with the deuterium-isotope shifts expected for protons in a mostly deuterated background (for example, as $-C^2H_2{}^1H$ in methyls). The deuterium shift can be large enough (around 0.04 ppm in alanine methyls[11]) to result in misinterpretation of NOEs as long-range. Of the 58 observable backbone amides in p7, at least 18 (31%) exhibit NOEs consistent with residual protonation (Fig. 2a). In addition, the observed cross-peaks tend to correlate with side-chain protons closer to the backbone. Alanines should be most susceptible to NOE artefacts from trace protonation since the methyls are close to the backbone amide proton and the three deuterium positions increase the probability of a proton being present. Indeed, six of eight assigned alanines show cross-peaks in the mixed-label sample that correlate with the intramolecular H_β chemical shift.

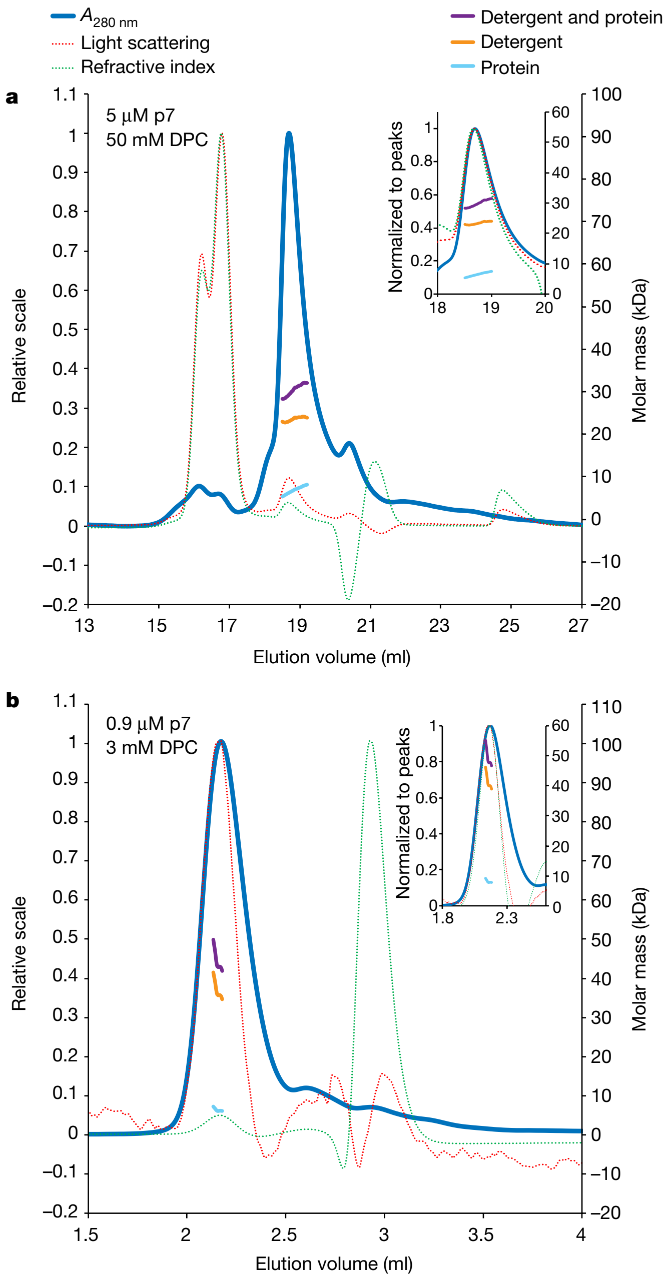


Fig. 1 | SEC–MALS of p7 in 3 mM and 50 mM DPC. a, SEC–MALS of a sample in conditions resembling those in which NMR was used to calculate the putative hexameric structure[2]. The protein was dissolved into 200 mM DPC and 6 M guanidine and reconstituted by dialysis as described[2,8], with a final DPC concentration of 50 mM. A DPC concentration of 50 mM was used[8] instead of 200 mM because the scattering intensity at 200 mM DPC saturates the detector[2]. A Superdex 200 10/300 column was used for size-exclusion chromatography. The calculated mass values were 6.4 ± 1.1 kDa for the protein and 23.5 ± 0.5 kDa for the associated detergent. Left axis shows A280 nm, light scattering and refractive index. Right axis shows molar masses calculated at each point for the protein, associated detergent and the detergent and protein complex. Insets show the three detector signals normalized at the p7 peak. The DPC concentration and maximum eluted monomeric p7 concentration (summit of peak) are indicated on the graphs. Reported masses denote the value at the peak summit, and the error is taken as the maximum difference from this value across the elution volume for which the molar mass is plotted. **b**, SEC–MALS of a sample prepared as for electron microscopy studies[2] (3 mM DPC), but at higher protein concentration. P7 was refolded in 200 mM DPC, then subjected to size-exclusion chromatography on a Superdex 200 10/300 column in 3 mM DPC to remove excess DPC. A small amount of protein aggregates with large scattering and refractive-index peaks was observed before the monomeric p7 peak (similar to those seen in **a**). The p7 monodisperse peak was collected and analysed by SEC–MALS with 3 mM DPC running buffer in a Superdex 200 5/150 column. The calculated molar masses were 8.4 ± 0.7 and 42.4 ± 3.7 kDa for protein and the associated detergent, respectively. All SEC–MALS conditions and calculated masses are summarized in Extended Data Table 1.

NOE cross-peaks assigned to intermolecular interactions with rimantadine in a ^{15}N–^{2}H-labelled sample containing 5 mM rimantadine are also consistent with trace protonation, since these peaks are present in the rimantadine-free mixed-label sample, and the cross-peaks correlate with intra-residue side-chain protons (Fig. 2b). Similar cross-peaks are also attributed to an amantadine interaction[2]. Among valine, leucine and isoleucine methyls, OuYang et al.[2] reported large methyl chemical-shift perturbations upon addition of rimantadine for Val7-γ2, Val25-γ2 and Val53-γ1 methyls, of which only Val25 and Val53 are near the proposed binding site. To provide greater data coverage, we used the backbone amide spectra of OuYang et al.[2] to calculate chemical-shift differences upon addition of 5 mM rimantadine. A large number of residues across the protein exhibit chemical-shift differences in an apparently nonspecific manner (Extended Data Fig. 2e), and residues identified as lining the rimantadine and amantadine binding site (figure 3c in ref. [2]) show some of the smallest chemical-shift differences. The distribution and magnitude of the shifts are similar to what was observed after addition of amantadine to a sample of monomeric p7 at pH 4.0 and 50 °C[12]. The K_d values for rimantadine reported by OuYang et al.[2] (13.2 μM and 63.6 μM) are at least three orders of magnitude higher than the equivalent values in membranes[13], and are consistent with nonspecific binding.

OuYang et al.[2] used residual dipolar couplings (RDCs) to calculate their p7 structure. Although RDCs do not directly report on oligomer stoichiometry, best fits of the RDC data to the deposited structures result in alignment tensors with large, non-zero rhombicities (larger than 0.48; average of 0.57) that are inconsistent with a tightly associated, symmetric oligomer[14]. We note also that refinement against a single set of amide-bond RDCs from one alignment medium will not yield a unique structure in the absence of information about long-range contacts[15]. This means that it is possible to fit the amide-bond RDCs to a p7-subunit structure that is influenced by incorrectly assigned intermolecular restraints.

In conclusion, we find that p7 is monomeric over a range of protein and DPC concentrations, including the NMR conditions used to determine an oligomeric structure[2], and that NOEs identified as unambiguously intermolecular are consistent with artefacts from residual protonation. We note that the protein concentration (200 nM monomer) used in the electron microscopy experiments cannot be analysed by NMR or SEC–MALS, and therefore oligomerization upon sample dilution for electron microscopy studies cannot be ruled out. Moreover, it cannot be excluded that the p7 oligomeric complexes observed by electron microscopy represent a small proportion of the total protein sample that is not detectable by NMR or SEC–MALS.

Methods

Estimation of complex size from NMR relaxation data. To exclude data from residues with internal motions faster or slower than the overall tumbling time, the rotational correlation time τ_c was calculated from the 20% trimmed means of the ^{15}N-relaxation rates[16], which were 1.34 s^{-1} and 14.80 s^{-1} for R_1 and R_2, respectively. ^{15}N-relaxation rates were measured using HSQC-based experiments. The molecular mass was calculated from τ_c using Stokes' law, assuming a hydration shell of

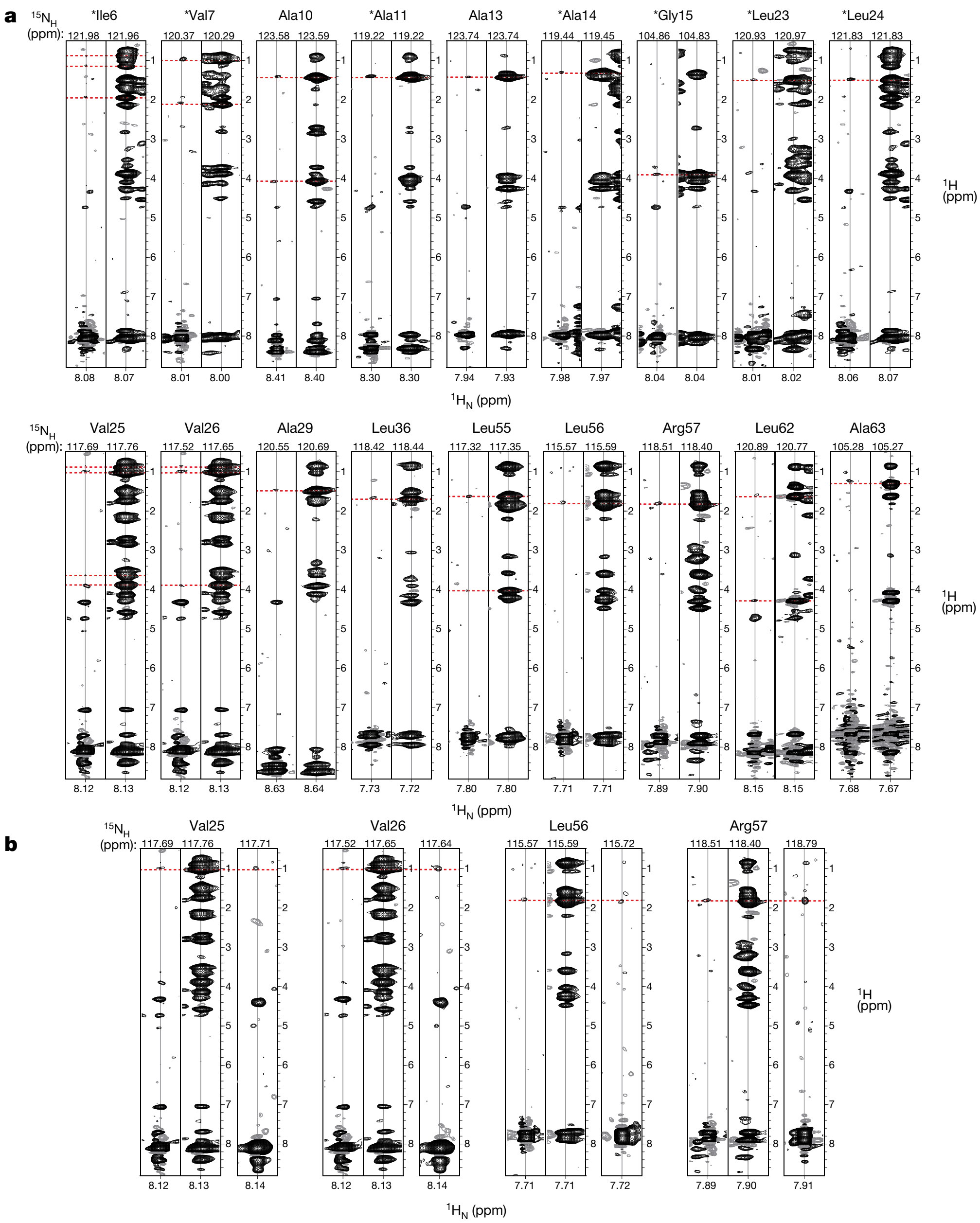


Fig. 2 | Evidence of residual protonation from comparison of 3D ^{15}N-edited NOESY strips. a, Alignments of indirect ^{1}H-dimension strips from the 3D ^{15}N-edited NOESY spectra of OuYang et al.[2]. For each residue, indicated at the top of the strips, the left strip is from the mixed-label sample (1:1 mixture of ^{15}N–^{2}H- and ^{13}C-labelled p7) and the right strip is from a ^{15}N–^{1}H-labelled sample. Positive contours are in black and negative contours in grey. Horizontal dashed red lines are added to show chemical-shift correlations between strips and are positioned at the chemical shifts of selected intramolecular NOEs in the fully protonated sample. Asterisks indicate strips from which OuYang et al.[2] identified putative intermolecular NOEs: Ile6 H_N (to Ile6 $H_{\gamma2}$), Val7 H_N (to Val5 $H_{\gamma1}$), Ala11 H_N (to Ala61 H_β), Ala14 H_N (to Ala63 H_β), Leu23 H_N (to Ala29 H_β) and Leu24 H_N (to Ala29 H_β). The positions of the putative intermolecular NOEs correlate well with positions of strong intra-residue peaks in the fully protonated sample: Ile6 H_N (to Ile6 $H_{\gamma2}$), Val7 H_N (to Val7 $H_{\gamma1}$), Ala11 H_N (to Ala11 H_β), Ala14 H_N (to Ala14 H_β), Leu23 H_N (to Leu23 H_β) and Leu24 H_N (to Leu24 H_β). Several strips from the mixed-label NOESY indicate more than one correlation to an intra-residue NOE in the fully protonated sample. For the putative intermolecular NOE at Gly15 H_N (to His59 $H_{\varepsilon1}$), no obvious, resolvable cross-peaks corresponding to His59 side-chain protons were identifiable in the mixed-label NOESY, however, trace protonation at the Gly15 Hα is apparent in the mixed-label NOESY strip. Diagrams showing the strip for this NOE were not presented in the supplementary information of OuYang et al.[2]. An eighth NOE, from Ala10 (H_N) to Ala61 (H_β), was not present in the restraint file deposited with the BMRB but was indicated in supplementary figure 8 of OuYang et al.[2]; it can also be explained as an intramolecular NOE arising from trace protonation. **b**, Analysis of putative intermolecular NOEs to rimantadine. For each residue, indicated at the top of the strips, the left and middle strips correspond to the NOESYs shown in **a**, and the additional strip on the right is from a ^{15}N–^{2}H-labelled sample containing 5 mM rimantadine. NOEs identified by OuYang et al. as arising from rimantadine (see also supplementary figure 6 in OuYang et al.[2]) correlate with intramolecular NOEs observed for the fully protonated sample and for the mixed-label sample, both of which have no rimantadine added.

1.5 water molecules and a solution viscosity of 0.702 centipoise at 37 °C. The different partial specific volumes of protein (0.73 $cm^3 g^{-1}$) and micellar DPC (0.94 $cm^3 g^{-1}$) were taken into account to calculate the number of attached detergent molecules.

SEC–MALS. SEC–MALS was performed on a Shimadzu Nexera HPLC, a MALS DAWN HELEOS II and a refractive index Optilab T-rEX detector (Wyatt Technology). Molar masses were determined with the protein conjugate analysis tool in Astra v.6.1.1.17 (Wyatt), except for detergent-alone samples for which the standard tool based on light scattering and refractive index was used. The d*n*/d*c* values used for analyses were 0.1398 ml g^{-1} for DPC (Anatrace), 0.1100 ml g^{-1} for SDS[17] and 0.185 ml g^{-1} for protein (standard value).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0561-9.

Benjamin P. Oestringer[1,2,5], Juan H. Bolivar[1,2], Mario Hensen[1,2], Jolyon K. Claridge[1,6,7], Chris Chipot[3,4], François Dehez[3], Nicole Holzmann[3], Nicole Zitzmann[1,2]* & Jason R. Schnell[1]*

[1]Department of Biochemistry, University of Oxford, Oxford, UK. [2]Oxford Glycobiology Institute, Department of Biochemistry, University of Oxford, Oxford, UK. [3]Laboratoire International Associé CNRS-University of Illinois at Urbana Champaign, UMR n°7019 Université de Lorraine, BP 70239, Vandoeuvre-lès-Nancy, France. [4]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [5]Present address: Immunocore Limited, Abingdon, UK. [6]Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium. [7]Structural and Molecular Microbiology, Structural Biology Research Center, VIB, Brussels, Belgium.
*e-mail: nicole.zitzmann@bioch.ox.ac.uk; jason.schnell@bioch.ox.ac.uk

Received: 11 September 2017; Accepted: 16 July 2018;
Published online 17 October 2018.

1. Madan, V. & Bartenschlager, R. Structural and functional properties of the hepatitis C virus p7 viroporin. *Viruses* **7**, 4461–4481 (2015).
2. OuYang, B. et al. Unusual architecture of the p7 channel from hepatitis C virus. *Nature* **498**, 521–525 (2013).
3. Chew, C. F., Vijayan, R., Chang, J., Zitzmann, N. & Biggin, P. C. Determination of pore-lining residues in the hepatitis C virus p7 protein. *Biophys. J.* **96**, L10–L12 (2009).
4. StGelais, C. et al. Determinants of hepatitis C virus p7 ion channel function and drug sensitivity identified in vitro. *J. Virol.* **83**, 7970–7981 (2009).
5. Luik, P. et al. The 3-dimensional structure of a hepatitis C virus p7 ion channel by electron microscopy. *Proc. Natl Acad. Sci. USA* **106**, 12712–12716 (2009).
6. Stansfeld, P. J. et al. MemProtMD: automated insertion of membrane protein structures into explicit lipid membranes. *Structure* **23**, 1350–1361 (2015).
7. Kalita, M. M., Griffin, S., Chou, J. J. & Fischer, W. B. Genotype-specific differences in structural features of hepatitis C virus (HCV) p7 membrane protein. *Biochim. Biophys. Acta* **1848**, 1383–1392 (2015).
8. Dev, J. & Brüschweiler, S. OuYang, B. & Chou, J. J. Transverse relaxation dispersion of the p7 membrane channel from hepatitis C virus reveals conformational breathing. *J. Biomol. NMR* **61**, 369–378 (2015).
9. Fernández, C. & Wider, G. TROSY in NMR studies of the structure and function of large biological macromolecules. *Curr. Opin. Struct. Biol.* **13**, 570–580 (2003).
10. Korepanova, A. & Matayoshi, E. D. HPLC–SEC characterization of membrane protein-detergent complexes. *Curr. Protoc. Protein Sci.* **Chapter 29**, 1–12 (2012).
11. Gardner, K. H., Rosen, M. K. & Kay, L. E. Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry* **36**, 1389–1401 (1997).
12. Cook, G. A., Dawson, L. A., Tian, Y. & Opella, S. J. Three-dimensional structure and interaction studies of hepatitis C virus p7 in 1,2-dihexanoyl-*sn*-glycero-3-phosphocholine by solution nuclear magnetic resonance. *Biochemistry* **52**, 5295–5303 (2013).
13. Breitinger, U., Farag, N. S., Ali, N. K. M. & Breitinger, H.-G. A. Patch–clamp study of hepatitis C p7 channels reveals genotype-specific sensitivity to inhibitors. *Biophys. J.* **110**, 2419–2429 (2016).
14. Al-Hashimi, H. M., Bolon, P. J. & Prestegard, J. H. Molecular symmetry as an aid to geometry determination in ligand protein complexes. *J. Magn. Reson.* **142**, 153–158 (2000).
15. Hus, J.-C. et al. 16-fold degeneracy of peptide plane orientations from residual dipolar couplings: analytical treatment and implications for protein structure determination. *J. Am. Chem. Soc.* **130**, 15927–15937 (2008).
16. Kay, L. E., Torchia, D. A. & Bax, A. Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry*. **28**, 15927–15937 (2008).
17. Tumolo, T., Angnes, L. & Baptista, M. S. Determination of the refractive index increment (d*n*/dc) of molecule and macromolecule solutions by surface plasmon resonance. *Anal. Biochem*. **333**, 273–279 (2004).
18. Turro, N. J. & Yekta, A. Luminescent probes for detergent solutions. A simple procedure for determination of the mean aggregation number of micelles. *J. Am. Chem. Soc.* **100**, 5951–5952 (1978).
19. Holzmann, N., Chipot, C., Penin, F. & Dehez, F. Assessing the physiological relevance of alternate architectures of the p7 protein of hepatitis C virus in different environments. *Bioorg. Med. Chem*. **24**, 4920–4927 (2016).
20. Slotboom, D. J., Duurkens, R. H., Olieman, K. & Erkens, G. B. Static light scattering to characterize membrane proteins in detergent solution. *Methods*. **46**, 73–82 (2008).

Author contributions B.P.O. performed protein expression, sample preparation, NMR experiments and data analysis, SEC–MALS experiments and analysis and helped write the paper; J.H.B. performed sample preparation, SEC–MALS experiments and analysis and helped write the paper; M.H. performed protein expression, sample preparation and SEC–MALS experiments; J.K.C. performed NMR experiments and data analysis; C.C., F.D. and N.H. performed molecular dynamics simulations; J.R.S. performed NMR experiments and reanalyzed NOESY spectra; N.Z. performed protein expression and sample preparation; J.R.S., N.Z., C.C. and F.D. conceived the study; and J.R.S. and N.Z. wrote the paper.

Competing interests Declared none.

Additional information

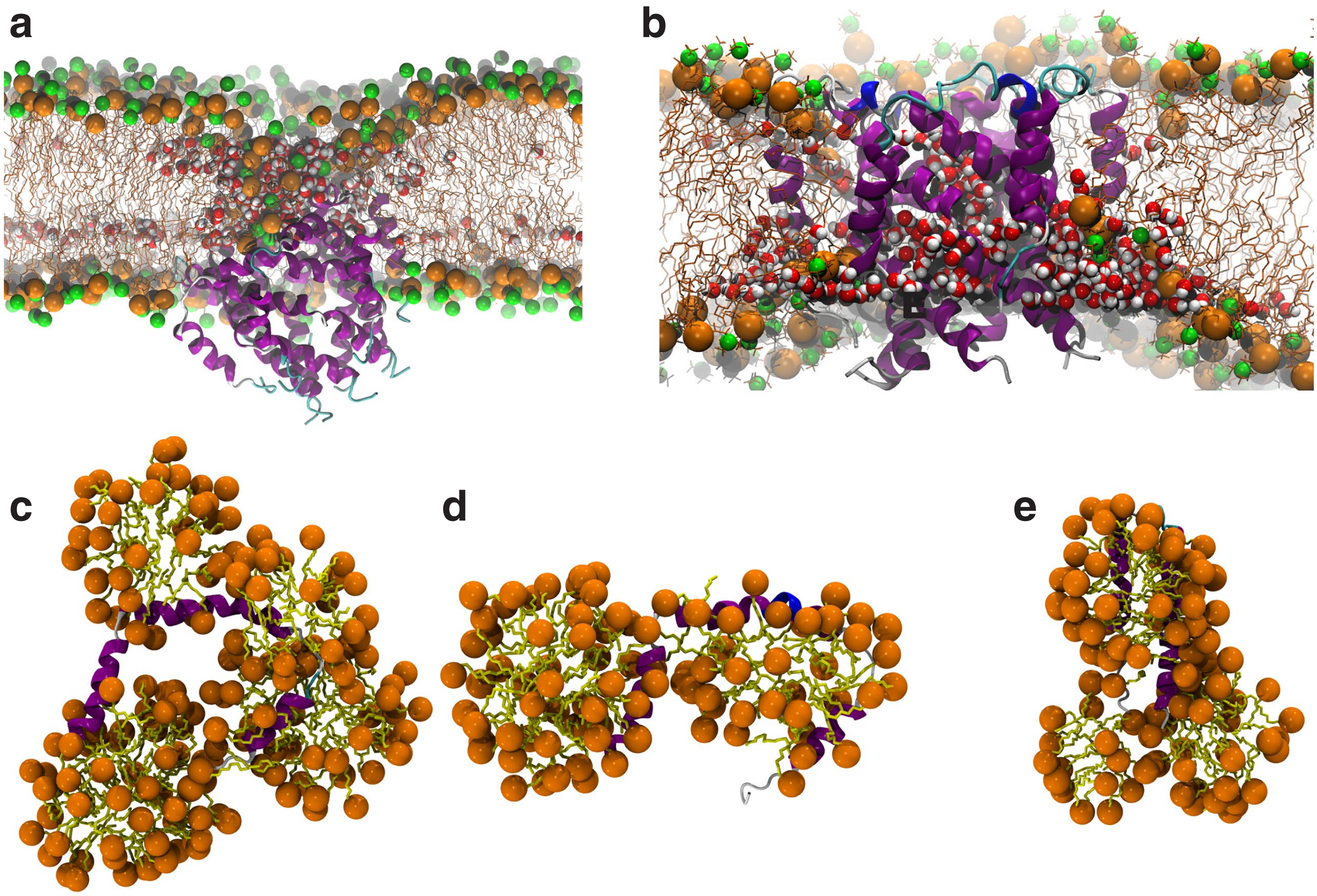
Extended data accompanies this Comment. https://doi.org/10.1038/s41586-018-0561-9

Supplementary information accompanies this Comment. https://doi.org/10.1038/s41586-018-0561-9

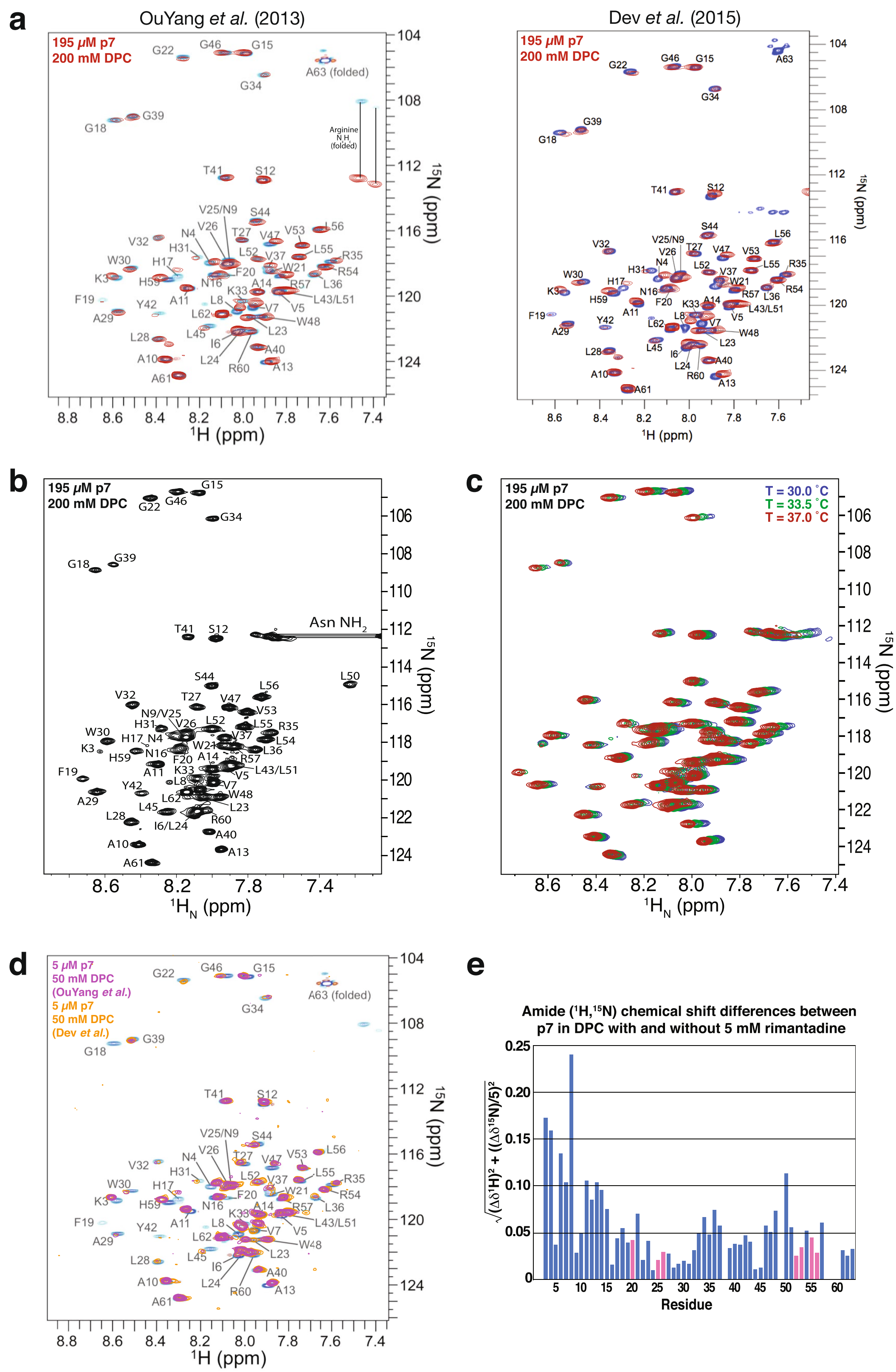
Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to N.Z. or J.R.S.

https://doi.org/10.1038/s41586-018-0561-9

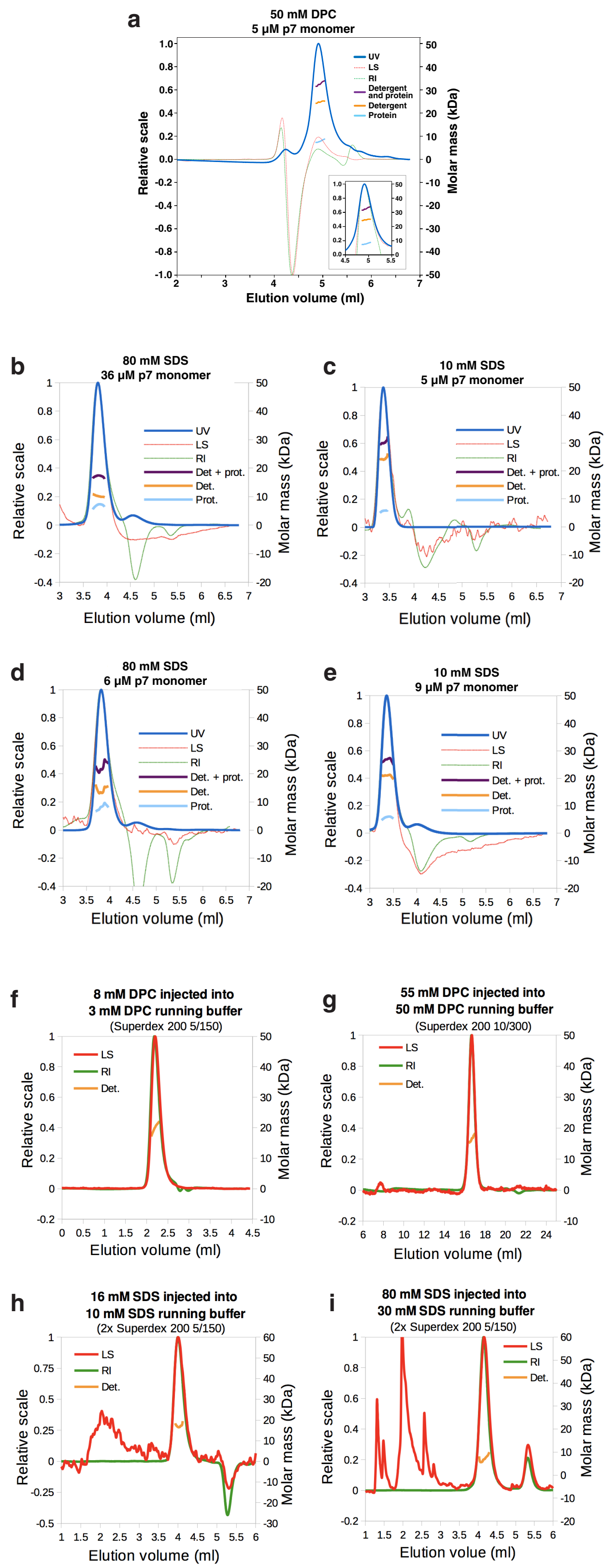


Extended Data Fig. 1 | Molecular dynamics of p7. a, **b**, Insertion of the proposed hexameric p7 structure[2] into lipid bilayers. MemProtMD[6] prediction for the hexamer insertion into a hydrated 1,2-dipalmitoyl-*sn*-glycero-3-phosphatidylcholine (DPPC) bilayer (**a**). The p7 structure[2] after insertion into a hydrated 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphatidylcholine (POPC) bilayer and simulated for 60 ns (**b**). Severe deformations and thinning defects of the bilayer can be seen, resulting in a large number of water molecules within the hydrophobic region of the bilayer. Water is shown as van der Waals spheres for oxygen (red) and hydrogen (white). For DPPC and POPC lipids, phosphorus and choline nitrogen positions are indicated with orange spheres and green spheres, respectively. In **b**, p7 α-helices and 3_{10}-helices are shown in magenta and blue, respectively. **c**–**e**, Simulations of monomeric p7 in 300 mM DPC at a protein to detergent ratio of 1:250. Independent 100-ns simulations of the horseshoe-like subunit conformation of the proposed hexameric p7 structure (**c**, **d**). At the end of the simulation, ~170 (**c**) and ~120 DPC (**d**) molecules, were observed bound to the protein. A hairpin conformation of p7 was simulated for 100 ns, at the end of which ~100 DPC molecules were observed bound to the protein (**e**). In **c**–**e**, p7 α-helices and 3_{10}-helices are shown in magenta and blue, respectively; the geometric centre of the DPC headgroup is indicated by an orange sphere, and the DPC hydrocarbon chain as yellow sticks. Only those DPC molecules bound to p7 are shown. The simulations in **b**–**e** were performed with the CHARMM36 all-atom force field using the protocol described[19].



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | NMR spectroscopy of p7 in DPC. a, Comparison of NMR spectra of p7, prepared as described[2], with previously published spectra. Left, 2D ^{1}H–^{15}N-TROSY spectrum of a sample of p7 in DPC produced in the present study (red cross peaks), recorded at 30 °C with a ^{1}H frequency of 600 MHz, overlaid with the spectrum published by OuYang et al.[2] (blue cross peaks). Right, our p7 spectrum (red cross peaks) overlaid with another published spectrum[8] (blue cross-peaks) to illustrate sample-to-sample variation. In both cases, the spectra were overlaid manually. **b**, ^{1}H–^{15}N HSQC of p7 in DPC at 37 °C. The spectrum was recorded at 600 MHz (^{1}H) on a Bruker spectrometer equipped with a CryoProbe. **c**, Temperature sensitivity of the spectrum of p7 in DPC. Spectra were recorded at three temperatures: 30 °C (blue), 33.5 °C (green) and 37 °C (red). The spectra have not been corrected for the temperature dependence of the ^{2}H-lock frequency. **d**, NMR spectra of p7 under SEC–MALS conditions. Spectra were recorded of 5 μM p7 with 50 mM DPC (detergent-to-protein ratio of 10,000) to match conditions at the SEC–MALS peak summit for a sample prepared as described[2] (violet) (see Fig. 1a and Extended Data Table 1), or as described by Dev et al.[8], with pre-gel filtration in 3 mM DPC buffer containing 100 mM NaCl (orange) (see Extended Data Fig. 3a). Before recording the NMR data, the sample in 3 mM DPC + 100 mM NaCl was dialysed against buffer (25 mM MES at pH 6.5, 3 mM DPC, and 1 mM DSS with 5% D_2O) overnight to remove salt, concentrated, and then diluted with NMR buffer (25 mM MES at pH 6.5, 50 mM DPC, and 1 mM DSS with 5% D_2O) to match the concentrations at the peak summit of the SEC–MALS elution (5 μM p7 and 50 mM DPC). Band-selective excitation short-transient transverse relaxation-optimized spectroscopy (BEST-TROSY) NMR experiments were recorded over ~20 h each on a 750-MHz spectrometer equipped with a CryoProbe to obtain sufficient signal-to-noise ratios for the dilute samples and overlaid with the published TROSY spectrum (supplementary figure 2a in OuYang et al.[2]). **e**, Backbone amide chemical-shift differences between a sample of p7 in DPC without and with 5 mM rimantadine. Data analysis was carried out on the spectra of OuYang et al.[2]. Chemical-shift differences were calculated as indicated on the vertical axis for the backbone amide resonances in the ^{15}N–^{2}H-labelled mixed-label sample and the ^{15}N–^{2}H-labelled sample containing 5 mM rimantadine. Residues in pink were previously identified[2] as forming the rimantadine-binding pocket (see figure 3c in ref. [2]). All reported concentrations are for monomeric protein.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | SEC–MALS of p7 and detergent alone. a, P7 in 50 mM DPC prepared as described by Dev. et al.[8]. The sample was run over two Superdex 200 5/150 columns in series in order to resolve the protein and micelle complex. The calculated masses were 7.8 ± 1.1 and 24.7 ± 0.6 kDa for protein and the associated detergent, respectively. The DPC concentration and maximum p7 monomeric concentration eluted (summit of peak) are indicated above the graph. **b**–**e**, SEC–MALS of p7 in 80 mM and 10 mM SDS. The SDS concentration and the maximum concentration of eluted monomeric p7 (summit of peak) are indicated above each graph. p7 is monomeric in SDS, as shown by its mobility on SDS–PAGE[2] (and our data, not shown). SEC–MALS molar mass analysis indicates, unambiguously, and in all cases, a monomeric state for p7 in SDS. The analysis also confirms the trend seen with DPC (Fig. 1 and **a** above), in that the amount of detergent associated with the protein is markedly and consistently higher for the samples at the lower detergent concentration (10 mM SDS) compared with samples in higher detergent concentration (80 mM SDS) (see data summary in Extended Data Table 1). Samples were run through two Superdex 200 5/150 columns connected in series for increased resolution. Sample injection volumes were 30 μl. Running buffer contained the indicated SDS concentrations and 10 mM sodium phosphate at pH 7.2. **f**–**i**, SEC–MALS of DPC and SDS micelles in the absence of protein. The detergent, its concentration in the injected sample and running buffers, and the chromatography columns used are indicated above each graph. As in **b**–**e**, SDS samples were run over two Superdex 200 5/150 in series to increase resolution. The $A_{280\,\mathrm{nm}}$ in samples without protein is negligible. Running buffers are the same as those used in experiments with protein (Fig. 1 and **a**–**e**), and the injected samples have higher detergent concentration to enable detection above the baseline. In **i**, the running buffer contained 30 mM SDS because the smaller size of micelles in 80 mM SDS in the absence of protein resulted in low signal-to-noise scattering. Molar masses (Det, orange line, right axis) increase slightly (by ~3 kDa for DPC and ~10 kDa for SDS) at concentrations close to the critical micelle concentration. Right axes: molar masses calculated at each point for protein (Prot; if protein is present), associated detergent (Det), and the detergent and protein complex (Det + Prot; if protein is present). Left axes: UV, $A_{280\,\mathrm{nm}}$; LS, light scattering; RI, refractive index. Detector signals are scaled to enable comparison. In **a**, the inset shows the detector signals normalized at the p7 peak. Extended Data Tables 1 and 2 show a summary of experimental conditions and mass calculations. The reported masses denote the value at the peak summit and the error is taken as the maximum difference from this value across the elution volume for which the molar mass is plotted. For samples with p7, negative and positive scattering and refractive-index peaks following the protein peaks are the result of distortions of the baseline upon sample injection that causes disequilibrium of detergent micelles in the running buffer[20].

Extended Data Table 1 | Summary of conditions and results for SEC–MALS studies of p7 in DPC and SDS

		Pre-injection			Eluted (peak summit)		Analysis of eluted p7 peak	
	Running buffer (mM)	p7 monomeric concentration (μM)	Det./Prot. (mol/mol)	Sample injection (μl)	p7 monomeric concentration (μM)	Det./Prot. (mol/mol)	Protein (kDa)	Detergent bound to p7 (kDa)
Fig. 1a	50 mM DPC	200	250	70	5	10000	6.4 ± 1.1	23.5 ± 0.5
Fig. 1b	3 mM DPC	15	200	15	0.9	3333	8.4 ± 0.7	42.4 ± 3.7
Ext. Data Fig. 3a	50 mM DPC	140	357	50	5	10000	7.8 ± 1.1	24.7 ± 0.6
Ext. Data Fig. 3b	80 mM SDS	464	172	30	36	2222	7.3 ± 1.8	10.2 ± 0.7
Ext. Data Fig. 3c	10 mM SDS	58	172	30	5	2000	5.9 ± 0.8	24.1 ± 2.3
Ext. Data Fig. 3d	80 mM SDS	58	1379	30	6	13333	8.3 ± 1.7	13.2 ± 3.1
Ext. Data Fig. 3e	10 mM SDS	120	83	30	9	1111	5.9 ± 1.2	21.0 ± 1.5

Data from Extended Data Fig. 3b, c show that at similar detergent/protein molar ratios the molar mass of detergent associated with p7 is approximately double in 10 mM SDS compared to in 80 mM SDS. Data from Extended Data Fig. 3d, e show that the result is consistent when changing the protein concentration. Together with the data for p7 in DPC (Fig. 1), the results suggest that at detergent concentrations close to the CMC (~8 mM for SDS and ~1.5 mM for DPC), a larger number of detergent molecules associate with p7. Running buffers contained the indicated detergent concentrations and 25 mM MES at pH 6.5 (DPC samples) or 10 mM sodium phosphate at pH 7.2 (SDS samples). The chromatography columns were Superdex 200 of sizes 10/300 (Fig. 1a) and 5/150 (Fig. 1b) or two 5/150 columns attached in series (Extended Data Fig. 3a–e), depending on sample volumes and concentrations, and resolution requirements. The samples (pre-injection) contained the same detergent concentration as the running buffer. Reported masses denote the value at the peak summit and the error is taken as the maximum difference from this value across the elution volume for which the molar mass is plotted in the SEC–MALS figures.

Extended Data Table 2 | Summary of conditions and results for SEC–MALS studies of DPC and SDS detergents in the absence of protein

Sample			Respective detergent concentration in running buffer (mM)	Micelle molar mass (kDa)	Superdex 200 dimensions
Detergent	Concentration (mM)	Volume (µl)			
DPC	8	5	3	20.2 ± 3.2	5/150
DPC	55	100	50	16.8 ± 1.8	10/300
SDS	16	30	10	17.2 ± 3.4	2x 5/150
SDS	80	30	30	6.9 ± 3.1	2x 5/150

Data are from experiments shown in Extended Data Fig. 3f–i. The micellar molar masses calculated from the SEC–MALS data at concentrations close to the CMC (~8 mM for SDS and ~1.5 mM for DPC) are in close agreement with those reported in the literature: ~19.0 kDa for DPC (Anatrace measurement in collaboration with M. Foster, University of Akron) and ~17.3 kDa for SDS[18]. At higher detergent concentrations, the micelle molar masses decrease by ~3 kDa for DPC and ~10 kDa for SDS. Running buffers contained the indicated detergent concentrations and 25 mM MES at pH 6.5 (DPC samples) or 10 mM sodium phosphate at pH 7.2 (SDS samples). Reported masses denote the value at the peak summit and the error is taken as the maximum difference from this value across the elution volume for which the molar mass is shown in the SEC–MALS figures.

Chen et al. reply

REPLYING TO B. P. Oestringer et al. *Nature* **562**, https://doi.org/10.1038/s41586-018-0561-9 (2018)

In the accompanying Comment[1], Oestringer et al. argue that the p7 protein from hepatitis C virus, which we used to determine the p7 oligomeric structure in dodecylphosphocholine (DPC) micelles[2], is monomeric. Here we show, with direct experimental evidence, that their assertion is wrong.

In our previous study[2], the hexameric state of p7 in DPC was indicated by two pieces of evidence. First, the negative-stain electron-microscopy analysis of the sample generated 2D reference-free class averages that are clearly hexameric. Second, the sample prepared in the same way, containing 1:1 mixture of ^{15}N–^{2}H-labelled monomer and ^{13}C-labelled monomer, showed unambiguous nuclear Overhauser effects (NOEs) between the amide protons of the deuterated monomers and the methyl protons of the fully-protonated monomers, indicating that the p7 reconstituted in DPC formed oligomers.

Oestringer et al.[1] argue that the ^{15}N–^{2}H-labelled p7 in the mixed sample was incompletely deuterated, and therefore the inter-monomer NOEs that we reported were actually intra-residue NOEs (figure 2 in ref. [1]). In our study, we recorded three ^{15}N-edited NOE spectroscopy (NOESY) spectra: one with a mixed sample containing 1:1 mixture of ^{15}N–^{2}H-labelled p7 and ^{13}C-labelled p7, and the other two with samples containing only ^{15}N–^{2}H-labelled p7 but in the presence of the drug amantadine or rimantadine. The latter two spectra are from samples that are not isotopically mixed, and can therefore be used as negative controls. The key question is whether or not the six most prominent inter-monomer NOEs reported (supplementary figure 8 in ref. [2]) are also present in the two negative controls. These NOEs relate to residues 7, 10, 11, 14, 23 and 24, but none of these are shown in figure 2b of Oestringer et al.[1]. We show that none of the six inter-monomer NOEs in the mixed sample are present in the negative control samples (Fig. 1), indicating that they have arisen from oligomerization of different isotopically labelled p7 monomers. We note that an unambiguous, complementary experiment to the mixed NOE experiment is measurement of mixed paramagnetic relaxation enhancement (PRE) using a sample in which half of the monomers are spin-labelled at the N terminus and the other half are ^{15}N-labelled. If the p7 in DPC forms oligomers, the NMR resonances of the relevant region should show strong PRE. Oestringer et al.[1] did not perform this simple and direct measurement to prove that the p7 is monomeric in DPC.

Oestringer et al.[1] presented data from NMR relaxation and size-exclusion chromatography coupled to multi-angle light scattering (SEC–MALS) studies to support their conclusion. NMR-derived rotational correlation time (T_c) is an ambiguous indicator of protein size owing to the presence of mobile regions. If T_c is to be used to infer the size of the p7 complex, then the mobile regions such as the first helix (H1) and the C-terminal residues (57–63) should be excluded. It has been shown that the average T_c at 30 °C excluding these regions is ~22 ns (supplementary figure 2 in ref. [3]), much larger than the 10.1 ns reported[1] by Oestringer et al. A T_c of 22 ns would be consistent with a 60–70 kDa complex. SEC–MALS analysis of small membrane peptides in detergent micelles is convoluted and unreliable, reflected by the inconsistency observed by Oestringer et al.[1] that the micelle sizes at different detergent concentrations are very different.

Oestringer et al.[1] also used other NMR-based arguments that are much less convincing. They argue that the molecular alignment tensor derived from residual dipolar coupling (RDC) measurements should be axially symmetric if p7 is a hexamer. However, this is only true if the p7 hexamer is completely rigid and symmetric. NMR relaxation data already indicates that large internal dynamics exist[3] and there is no a priori reason to suppose that these motions are symmetrical.

In conclusion, Oestringer et al.[1] present multiple indirect arguments that p7 is monomeric in DPC, but fail to convincingly refute the obvious inter-monomer NOEs, which represent direct evidence of oligomeric assembly. Their conclusion is further challenged by a recent study showing that the oligomeric assembly of p7 in DPC is consistent with that in bicelles that closely mimic the membrane environment[3].

Wen Chen, Bo OuYang and James J. Chou are solely responsible for this Reply; other authors from the original Letter who were involved in structural and functional studies of p7 did not contribute to this response and are not listed here. Readers are welcome to contact the authors for further unpublished data including intermolecular NOE and PRE data.

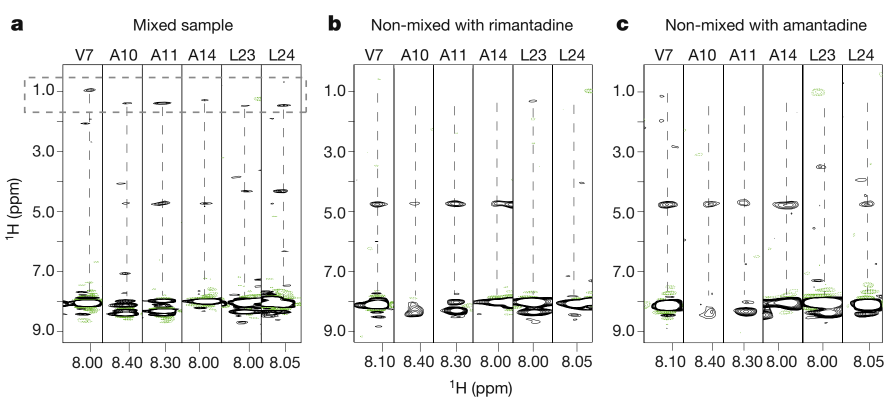


Fig. 1 | Mixed NOEs and negative controls. a, NOE strips from the mixed sample containing 50% ^{15}N–^{2}H-labelled p7 (5a) and 50% ^{13}C-labelled p7(5a). Inter-protomer NOEs are indicated in the dashed box. **b**, **c**, NOE strips from non-mixed samples containing 100% ^{15}N–^{2}H-labelled p7(5a) and the drug rimantadine and amantadine, respectively.

BRIEF COMMUNICATIONS ARISING

Wen Chen[1], Bo OuYang[2] & James J. Chou[1]*
[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA. [2]State Key Laboratory of Molecular Biology, National Center for Protein Science Shanghai, Shanghai Science Research Center, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai, China. *e-mail: james_chou@hms.harvard.edu

Published online 17 October 2018.

1. Oestringer, B. P. et al. Re-evaluating the p7 viroporin structure. *Nature* **562**, http://doi.org/10.1038/s41586-018-0561-9 (2018).
2. OuYang, B. et al. Unusual architecture of the p7 channel from hepatitis C virus. *Nature* **498**, 521–525 (2013).
3. Chen, W. et al. The unusual transmembrane partition of the hexameric channel of the hepatitis C virus. *Structure* **26**, 627–634 (2018).

Author contributions W.C., B.O. and J.J.C. wrote the response.

Competing interests Declared none.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.J.C.

https://doi.org/10.1038/s41586-018-0562-8

Re-evaluating the p7 viroporin structure

ARISING FROM B. OuYang et al. *Nature* **498**, 521–525 (2013); https://doi.org/10.1038/nature12283

The hepatitis C virus (HCV) p7 viroporin is a membrane protein required for virus propagation in vivo that assembles into hexamers and heptamers in membranes, exhibits ion channel activity, and is an attractive target against HCV infection[1]. OuYang and colleagues reported an oligomeric structure of p7 solubilized in dodecylphosphocholine (DPC) detergent, with unexpected features[2]. Here we show that p7 is monomeric in the conditions that were used to determine its oligomeric structure and that the data presented as evidence for intermolecular contacts is likely to arise from incomplete protein deuteration. We conclude that p7 is monomeric under NMR conditions, and that the oligomeric structure proposed by OuYang et al.[2] is artefactual. There is a Reply to this Comment by Chen, W. et al. *Nature* **562**, https://doi.org/10.1038/s41586-018-0562-8 (2018).

Unexpected features of the p7 oligomeric structure[2] include: (i) the His17 side chain, known to be involved in ion conduction[3,4], points outward towards the membrane bilayer; (ii) the orientation of the best fit of the oligomeric structure to the electron microscopy envelope contradicts antibody binding data[5]; and (iii) a short outer transmembrane helix exposes polar residues to the hydrophobic region of the membrane so that the structure cannot be accommodated in lipid bilayers without large structural rearrangements or membrane thinning[6,7] (Extended Data Fig. 1a, b).

We expressed, purified, and reconstituted into dodecylphosphocholine (DPC) ^{15}N-labelled protein corresponding to p7 of the genotype 5a isolate EUH1480, which contains five amino acid substitutions (p7(5a)[2], hereafter referred to as p7). Overlays of backbone ^{1}H–^{15}N-correlation NMR spectra confirm that the solution conditions and protein conformation are similar to those studied previously[2,8] (Extended Data Fig. 2a). High-quality backbone spectra of fully protonated p7 in DPC could be recorded at 37 °C using conventional, heteronuclear single quantum correlation (HSQC)-based experiments (Extended Data Fig. 2b), which is unexpected for a protein complex of 60–80 kDa[9] (the molecular mass of the protein oligomer and the associated detergent micelle). The spectra were unchanged between 30 °C and 37 °C, indicating that the protein adopted a similar conformation over this temperature range (Extended Data Fig. 2c).

Two methods were used to obtain information on the size of the p7–detergent-micelle complex. First, we estimated the effective protein mass using the rotational correlation time derived from ^{15}N R_1 and R_2 relaxation rates. The rotational correlation time, 10.1 ns at 37 °C, corresponds to 39.3 kDa, similar to a hexameric p7 (approximately 41 kDa) in the absence of a detergent micelle. Alternatively, the same rotational correlation time corresponds to monomeric p7 and a micelle of around 70 detergent molecules. We then applied size-exclusion chromatography coupled to multi-angle light scattering (SEC–MALS), which decomposes masses into protein and detergent contributions, and is reliable for proteins such as p7, which has a high molar extinction coefficient (17,990 M^{-1} cm^{-1}) and a relatively high light scattering intensity owing to the detergent micelle[10]. P7 samples reconstituted as described[2] and analysed using SEC–MALS in conditions similar to those used for NMR (50 mM DPC; Fig. 1a) or electron microscopy studies (3 mM DPC but a higher protein concentration; Fig. 1b) indicated protein masses consistent with monomers (6.4 ± 1.1 kDa and 8.4 ± 0.7 kDa, respectively) associated with detergent micelles of 23.5 ± 0.5 kDa and 42.4 ± 3.7 kDa, respectively (Extended Data Table 1). A third sample prepared using a different protocol[8] (50 mM DPC), indicated a protein mass of 7.8 ± 1.1 kDa associated with a detergent micelle of 24.7 ± 0.6 kDa (Extended Data Fig. 3a). Moreover, NMR spectra matching exactly the protein and detergent concentrations at the monomeric SEC–MALS peak summit (5 μM protein and 50 mM DPC) show good overlap with the spectrum published in ref. [2] (Extended Data Fig. 2d).

The SEC–MALS results are confirmed by comparison with p7 solubilized in sodium dodecyl sulfate (SDS) at 'low' (less than two times the critical micelle concentration (CMC)) and 'high' (at least ten times the CMC) detergent concentrations. SDS is strongly denaturing, and p7 runs as a monomer on SDS–PAGE[2]. Results for p7 in high (80 mM) and low (10 mM) SDS concentrations are consistent with a monomer, as expected, and similar to the results with DPC, in that low concentrations of SDS lead to an increase in detergent-micelle size with no change in oligomerization state (Extended Data Fig 3b–e and Extended Data Table 1). The same trend in micelle size with detergent concentration, albeit less marked, is observed in the absence of protein (Extended Data Fig. 3f–i and Extended Data Table 2).

OuYang et al.[2] identified putative intermolecular nuclear Overhauser effects (NOEs) from a 3D ^{15}N-edited NOE spectroscopy with transverse relaxation-optimized spectroscopy (NOESY–TROSY) experiment, without filtering for ^{13}C-attached protons, on a sample containing a 1:1 mixture of ^{15}N–^{2}H- and ^{13}C-labelled proteins. This experiment can provide intermolecular NOEs between exchangeable amide protons of the ^{15}N–^{2}H-labelled protein and non-exchangeable aliphatic protons of the ^{13}C-labelled protein. Artefactual cross-peaks can arise from incomplete deuteration of the ^{15}N–^{2}H-labelled protein, and a control NOESY experiment must be recorded on an unmixed but otherwise identical sample of ^{15}N–^{2}H-labelled protein, since commercial sources of D_2O and deuterated glucose contain at least small amounts of protons and additional protons can be carried over from starter cultures and humidity in the air. OuYang et al.[2] did not report control experiments, but comparison of the mixed-label NOESY spectra with a NOESY spectrum collected on fully protonated protein can indicate whether the NOEs are consistent with trace protonation of the ^{15}N–^{2}H-labelled protein. The NOESY spectra were aligned in the indirect ^{1}H dimensions such that the amide-proton chemical shifts, which exhibit relatively small deuterium-isotope shifts, were consistent. Seven NOEs were identified by OuYang et al.[2] as unambiguously intermolecular, and for each of these that could be identified in their mixed-label NOESY spectrum, there is a corresponding strong peak in their fully protonated sample NOESY close to the position expected for an intra-residue proton (Fig. 2a). The mixed-label NOEs were slightly shifted up-field on the order of 0.01 ppm, consistent with the deuterium-isotope shifts expected for protons in a mostly deuterated background (for example, as –$C^{2}H_{2}{}^{1}H$ in methyls). The deuterium shift can be large enough (around 0.04 ppm in alanine methyls[11]) to result in misinterpretation of NOEs as long-range. Of the 58 observable backbone amides in p7, at least 18 (31%) exhibit NOEs consistent with residual protonation (Fig. 2a). In addition, the observed cross-peaks tend to correlate with side-chain protons closer to the backbone. Alanines should be most susceptible to NOE artefacts from trace protonation since the methyls are close to the backbone amide proton and the three deuterium positions increase the probability of a proton being present. Indeed, six of eight assigned alanines show cross-peaks in the mixed-label sample that correlate with the intramolecular H_β chemical shift.

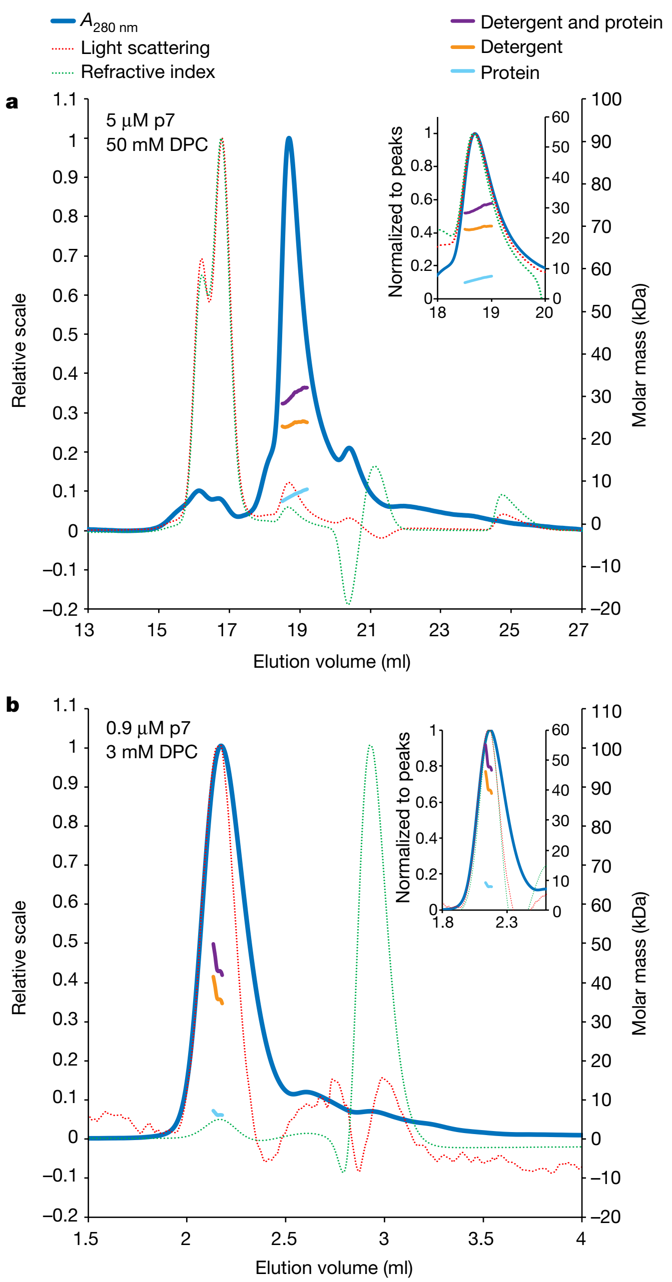


Fig. 1 | SEC–MALS of p7 in 3 mM and 50 mM DPC. a, SEC–MALS of a sample in conditions resembling those in which NMR was used to calculate the putative hexameric structure[2]. The protein was dissolved into 200 mM DPC and 6 M guanidine and reconstituted by dialysis as described[2,8], with a final DPC concentration of 50 mM. A DPC concentration of 50 mM was used[8] instead of 200 mM because the scattering intensity at 200 mM DPC saturates the detector[2]. A Superdex 200 10/300 column was used for size-exclusion chromatography. The calculated mass values were 6.4 ± 1.1 kDa for the protein and 23.5 ± 0.5 kDa for the associated detergent. Left axis shows A280 nm, light scattering and refractive index. Right axis shows molar masses calculated at each point for the protein, associated detergent and the detergent and protein complex. Insets show the three detector signals normalized at the p7 peak. The DPC concentration and maximum eluted monomeric p7 concentration (summit of peak) are indicated on the graphs. Reported masses denote the value at the peak summit, and the error is taken as the maximum difference from this value across the elution volume for which the molar mass is plotted. **b**, SEC–MALS of a sample prepared as for electron microscopy studies[2] (3 mM DPC), but at higher protein concentration. P7 was refolded in 200 mM DPC, then subjected to size-exclusion chromatography on a Superdex 200 10/300 column in 3 mM DPC to remove excess DPC. A small amount of protein aggregates with large scattering and refractive-index peaks was observed before the monomeric p7 peak (similar to those seen in **a**). The p7 monodisperse peak was collected and analysed by SEC–MALS with 3 mM DPC running buffer in a Superdex 200 5/150 column. The calculated molar masses were 8.4 ± 0.7 and 42.4 ± 3.7 kDa for protein and the associated detergent, respectively. All SEC–MALS conditions and calculated masses are summarized in Extended Data Table 1.

NOE cross-peaks assigned to intermolecular interactions with rimantadine in a ^{15}N–^{2}H-labelled sample containing 5 mM rimantadine are also consistent with trace protonation, since these peaks are present in the rimantadine-free mixed-label sample, and the cross-peaks correlate with intra-residue side-chain protons (Fig. 2b). Similar cross-peaks are also attributed to an amantadine interaction[2]. Among valine, leucine and isoleucine methyls, OuYang et al.[2] reported large methyl chemical-shift perturbations upon addition of rimantadine for Val7-γ2, Val25-γ2 and Val53-γ1 methyls, of which only Val25 and Val53 are near the proposed binding site. To provide greater data coverage, we used the backbone amide spectra of OuYang et al.[2] to calculate chemical-shift differences upon addition of 5 mM rimantadine. A large number of residues across the protein exhibit chemical-shift differences in an apparently nonspecific manner (Extended Data Fig. 2e), and residues identified as lining the rimantadine and amantadine binding site (figure 3c in ref. [2]) show some of the smallest chemical-shift differences. The distribution and magnitude of the shifts are similar to what was observed after addition of amantadine to a sample of monomeric p7 at pH 4.0 and 50 °C[12]. The K_d values for rimantadine reported by OuYang et al.[2] (13.2 μM and 63.6 μM) are at least three orders of magnitude higher than the equivalent values in membranes[13], and are consistent with nonspecific binding.

OuYang et al.[2] used residual dipolar couplings (RDCs) to calculate their p7 structure. Although RDCs do not directly report on oligomer stoichiometry, best fits of the RDC data to the deposited structures result in alignment tensors with large, non-zero rhombicities (larger than 0.48; average of 0.57) that are inconsistent with a tightly associated, symmetric oligomer[14]. We note also that refinement against a single set of amide-bond RDCs from one alignment medium will not yield a unique structure in the absence of information about long-range contacts[15]. This means that it is possible to fit the amide-bond RDCs to a p7-subunit structure that is influenced by incorrectly assigned intermolecular restraints.

In conclusion, we find that p7 is monomeric over a range of protein and DPC concentrations, including the NMR conditions used to determine an oligomeric structure[2], and that NOEs identified as unambiguously intermolecular are consistent with artefacts from residual protonation. We note that the protein concentration (200 nM monomer) used in the electron microscopy experiments cannot be analysed by NMR or SEC–MALS, and therefore oligomerization upon sample dilution for electron microscopy studies cannot be ruled out. Moreover, it cannot be excluded that the p7 oligomeric complexes observed by electron microscopy represent a small proportion of the total protein sample that is not detectable by NMR or SEC–MALS.

Methods

Estimation of complex size from NMR relaxation data. To exclude data from residues with internal motions faster or slower than the overall tumbling time, the rotational correlation time τ_c was calculated from the 20% trimmed means of the ^{15}N-relaxation rates[16], which were 1.34 s^{-1} and 14.80 s^{-1} for R_1 and R_2, respectively. ^{15}N-relaxation rates were measured using HSQC-based experiments. The molecular mass was calculated from τ_c using Stokes' law, assuming a hydration shell of

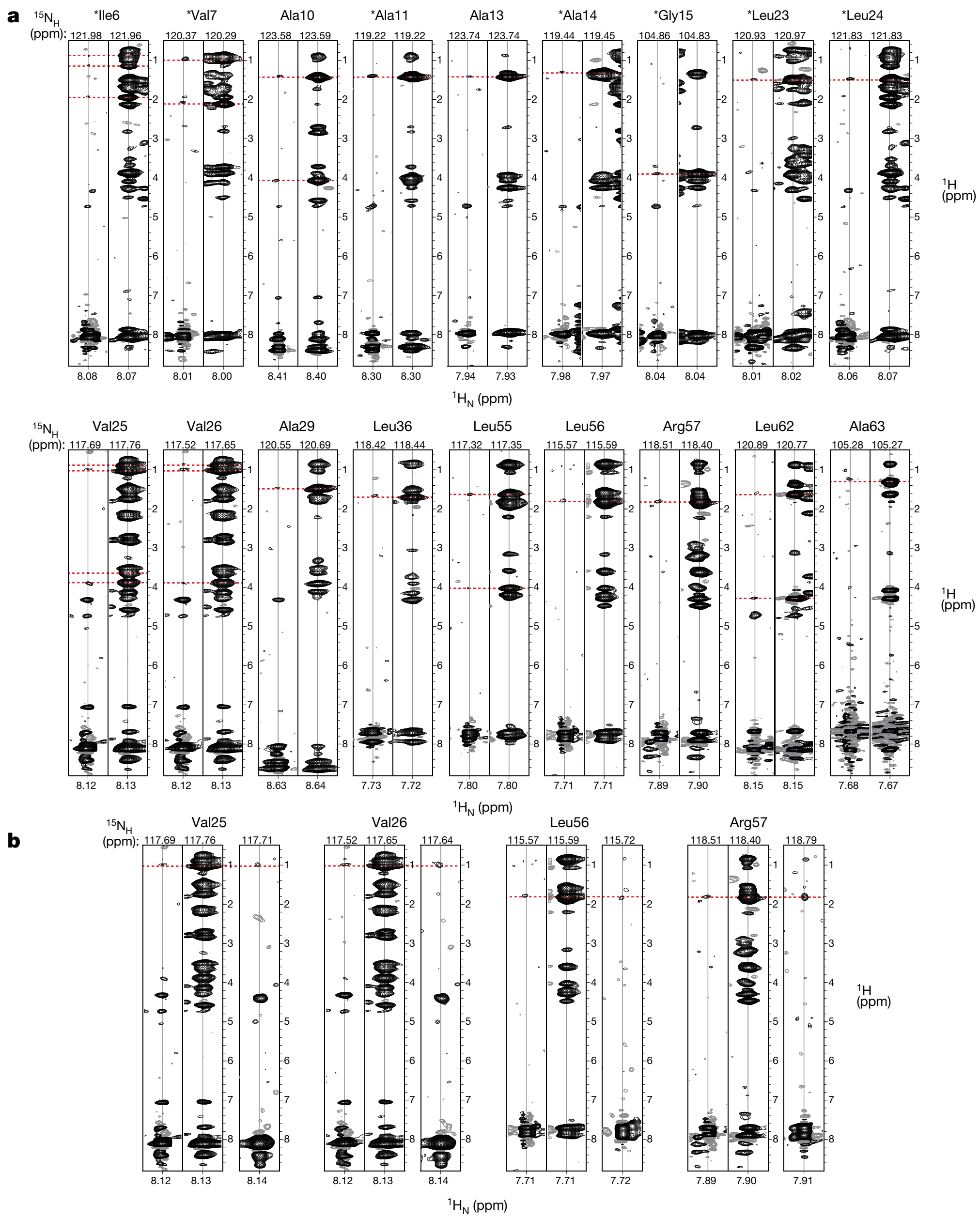


Fig. 2 | Evidence of residual protonation from comparison of 3D ^{15}N-edited NOESY strips. a, Alignments of indirect ^{1}H-dimension strips from the 3D ^{15}N-edited NOESY spectra of OuYang et al.[2]. For each residue, indicated at the top of the strips, the left strip is from the mixed-label sample (1:1 mixture of ^{15}N–^{2}H- and ^{13}C-labelled p7) and the right strip is from a ^{15}N–^{1}H-labelled sample. Positive contours are in black and negative contours in grey. Horizontal dashed red lines are added to show chemical-shift correlations between strips and are positioned at the chemical shifts of selected intramolecular NOEs in the fully protonated sample. Asterisks indicate strips from which OuYang et al.[2] identified putative intermolecular NOEs: Ile6 H_N (to Ile6 $H_{\gamma 2}$), Val7 H_N (to Val5 $H_{\gamma 1}$), Ala11 H_N (to Ala61 H_β), Ala14 H_N (to Ala63 H_β), Leu23 H_N (to Ala29 H_β) and Leu24 H_N (to Ala29 H_β). The positions of the putative intermolecular NOEs correlate well with positions of strong intra-residue peaks in the fully protonated sample: Ile6 H_N (to Ile6 $H_{\gamma 2}$), Val7 H_N (to Val7 $H_{\gamma 1}$), Ala11 H_N (to Ala11 H_β), Ala14 H_N (to Ala14 H_β), Leu23 H_N (to Leu23 H_β) and Leu24 H_N (to Leu24 H_β). Several strips from the mixed-label NOESY indicate more than one correlation to an intra-residue NOE in the fully protonated sample. For the putative intermolecular NOE at Gly15 H_N (to His59 $H_{\varepsilon 1}$), no obvious, resolvable cross-peaks corresponding to His59 side-chain protons were identifiable in the mixed-label NOESY, however, trace protonation at the Gly15 Hα is apparent in the mixed-label NOESY strip. Diagrams showing the strip for this NOE were not presented in the supplementary information of OuYang et al.[2]. An eighth NOE, from Ala10 (H_N) to Ala61 (H_β), was not present in the restraint file deposited with the BMRB but was indicated in supplementary figure 8 of OuYang et al.[2]; it can also be explained as an intramolecular NOE arising from trace protonation. **b**, Analysis of putative intermolecular NOEs to rimantadine. For each residue, indicated at the top of the strips, the left and middle strips correspond to the NOESYs shown in **a**, and the additional strip on the right is from a ^{15}N–^{2}H-labelled sample containing 5 mM rimantadine. NOEs identified by OuYang et al. as arising from rimantadine (see also supplementary figure 6 in OuYang et al.[2]) correlate with intramolecular NOEs observed for the fully protonated sample and for the mixed-label sample, both of which have no rimantadine added.

1.5 water molecules and a solution viscosity of 0.702 centipoise at 37 °C. The different partial specific volumes of protein (0.73 $cm^3 g^{-1}$) and micellar DPC (0.94 $cm^3 g^{-1}$) were taken into account to calculate the number of attached detergent molecules.

SEC–MALS. SEC–MALS was performed on a Shimadzu Nexera HPLC, a MALS DAWN HELEOS II and a refractive index Optilab T-rEX detector (Wyatt Technology). Molar masses were determined with the protein conjugate analysis tool in Astra v.6.1.1.17 (Wyatt), except for detergent-alone samples for which the standard tool based on light scattering and refractive index was used. The dn/dc values used for analyses were 0.1398 ml g^{-1} for DPC (Anatrace), 0.1100 ml g^{-1} for SDS[17] and 0.185 ml g^{-1} for protein (standard value).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0561-9.

Benjamin P. Oestringer[1,2,5], Juan H. Bolivar[1,2], Mario Hensen[1,2], Jolyon K. Claridge[1,6,7], Chris Chipot[3,4], François Dehez[3], Nicole Holzmann[3], Nicole Zitzmann[1,2]* & Jason R. Schnell[1]*

[1]Department of Biochemistry, University of Oxford, Oxford, UK. [2]Oxford Glycobiology Institute, Department of Biochemistry, University of Oxford, Oxford, UK. [3]Laboratoire International Associé CNRS-University of Illinois at Urbana Champaign, UMR n°7019 Université de Lorraine, BP 70239, Vandoeuvre-lès-Nancy, France. [4]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [5]Present address: Immunocore Limited, Abingdon, UK. [6]Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium. [7]Structural and Molecular Microbiology, Structural Biology Research Center, VIB, Brussels, Belgium.
*e-mail: nicole.zitzmann@bioch.ox.ac.uk; jason.schnell@bioch.ox.ac.uk

Received: 11 September 2017; Accepted: 16 July 2018;
Published online 17 October 2018.

1. Madan, V. & Bartenschlager, R. Structural and functional properties of the hepatitis C virus p7 viroporin. *Viruses* **7**, 4461–4481 (2015).
2. OuYang, B. et al. Unusual architecture of the p7 channel from hepatitis C virus. *Nature* **498**, 521–525 (2013).
3. Chew, C. F., Vijayan, R., Chang, J., Zitzmann, N. & Biggin, P. C. Determination of pore-lining residues in the hepatitis C virus p7 protein. *Biophys. J.* **96**, L10–L12 (2009).
4. StGelais, C. et al. Determinants of hepatitis C virus p7 ion channel function and drug sensitivity identified in vitro. *J. Virol.* **83**, 7970–7981 (2009).
5. Luik, P. et al. The 3-dimensional structure of a hepatitis C virus p7 ion channel by electron microscopy. *Proc. Natl Acad. Sci. USA* **106**, 12712–12716 (2009).
6. Stansfeld, P. J. et al. MemProtMD: automated insertion of membrane protein structures into explicit lipid membranes. *Structure* **23**, 1350–1361 (2015).
7. Kalita, M. M., Griffin, S., Chou, J. J. & Fischer, W. B. Genotype-specific differences in structural features of hepatitis C virus (HCV) p7 membrane protein. *Biochim. Biophys. Acta* **1848**, 1383–1392 (2015).
8. Dev, J. & Brüschweiler, S. OuYang, B. & Chou, J. J. Transverse relaxation dispersion of the p7 membrane channel from hepatitis C virus reveals conformational breathing. *J. Biomol. NMR* **61**, 369–378 (2015).
9. Fernández, C. & Wider, G. TROSY in NMR studies of the structure and function of large biological macromolecules. *Curr. Opin. Struct. Biol.* **13**, 570–580 (2003).
10. Korepanova, A. & Matayoshi, E. D. HPLC–SEC characterization of membrane protein-detergent complexes. *Curr. Protoc. Protein Sci.* **Chapter 29**, 1–12 (2012).
11. Gardner, K. H., Rosen, M. K. & Kay, L. E. Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry* **36**, 1389–1401 (1997).
12. Cook, G. A., Dawson, L. A., Tian, Y. & Opella, S. J. Three-dimensional structure and interaction studies of hepatitis C virus p7 in 1,2-dihexanoyl-*sn*-glycero-3-phosphocholine by solution nuclear magnetic resonance. *Biochemistry* **52**, 5295–5303 (2013).
13. Breitinger, U., Farag, N. S., Ali, N. K. M. & Breitinger, H.-G. A. Patch–clamp study of hepatitis C p7 channels reveals genotype-specific sensitivity to inhibitors. *Biophys. J.* **110**, 2419–2429 (2016).
14. Al-Hashimi, H. M., Bolon, P. J. & Prestegard, J. H. Molecular symmetry as an aid to geometry determination in ligand protein complexes. *J. Magn. Reson.* **142**, 153–158 (2000).
15. Hus, J.-C. et al. 16-fold degeneracy of peptide plane orientations from residual dipolar couplings: analytical treatment and implications for protein structure determination. *J. Am. Chem. Soc.* **130**, 15927–15937 (2008).
16. Kay, L. E., Torchia, D. A. & Bax, A. Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry*. **28**, 15927–15937 (2008).
17. Tumolo, T., Angnes, L. & Baptista, M. S. Determination of the refractive index increment (d*n*/dc) of molecule and macromolecule solutions by surface plasmon resonance. *Anal. Biochem*. **333**, 273–279 (2004).
18. Turro, N. J. & Yekta, A. Luminescent probes for detergent solutions. A simple procedure for determination of the mean aggregation number of micelles. *J. Am. Chem. Soc.* **100**, 5951–5952 (1978).
19. Holzmann, N., Chipot, C., Penin, F. & Dehez, F. Assessing the physiological relevance of alternate architectures of the p7 protein of hepatitis C virus in different environments. *Bioorg. Med. Chem*. **24**, 4920–4927 (2016).
20. Slotboom, D. J., Duurkens, R. H., Olieman, K. & Erkens, G. B. Static light scattering to characterize membrane proteins in detergent solution. *Methods*. **46**, 73–82 (2008).

Author contributions B.P.O. performed protein expression, sample preparation, NMR experiments and data analysis, SEC–MALS experiments and analysis and helped write the paper; J.H.B. performed sample preparation, SEC–MALS experiments and analysis and helped write the paper; M.H. performed protein expression, sample preparation and SEC–MALS experiments; J.K.C. performed NMR experiments and data analysis; C.C., F.D. and N.H. performed molecular dynamics simulations; J.R.S. performed NMR experiments and reanalyzed NOESY spectra; N.Z. performed protein expression and sample preparation; J.R.S., N.Z., C.C. and F.D. conceived the study; and J.R.S. and N.Z. wrote the paper.

Competing interests Declared none.

Additional information

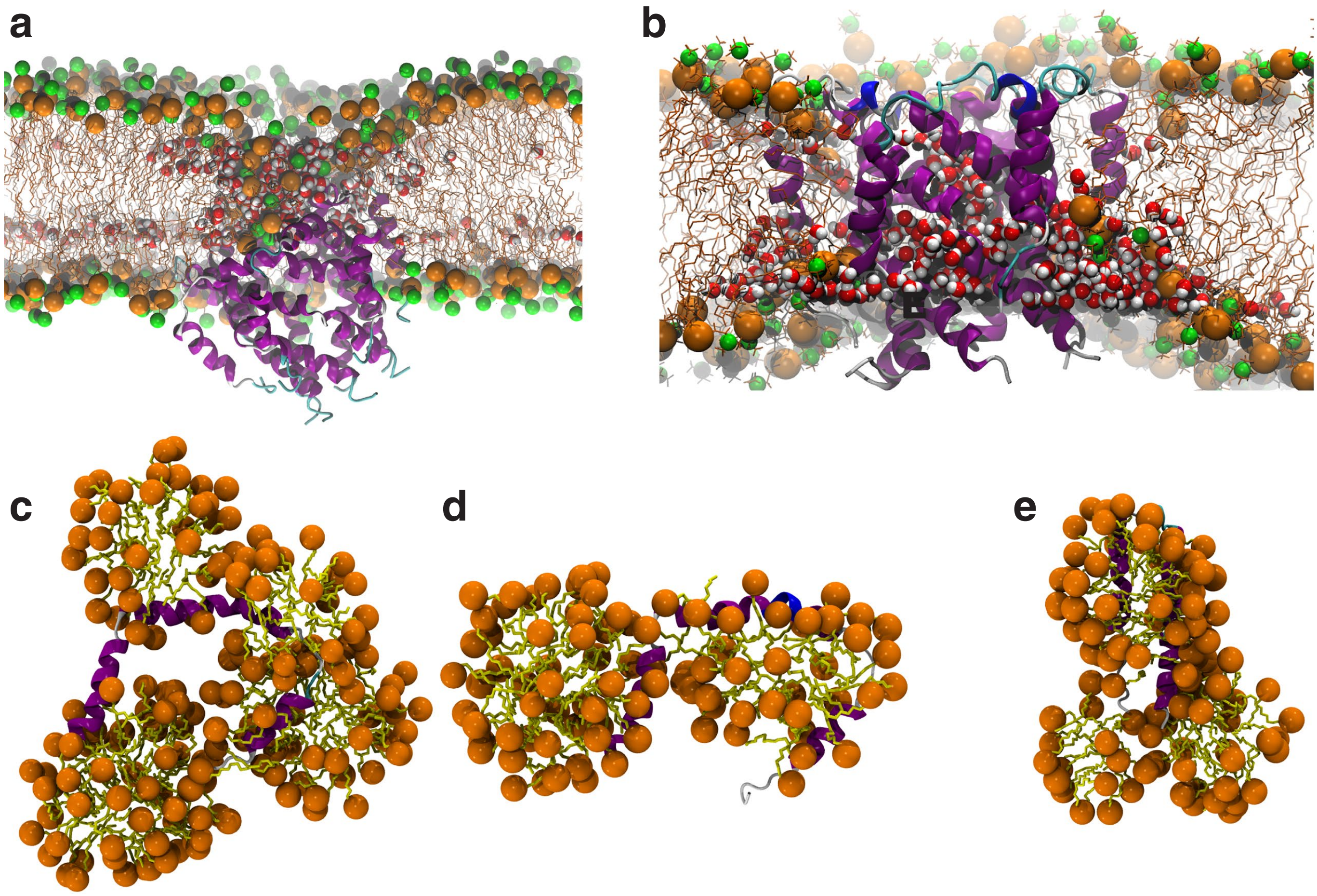
Extended data accompanies this Comment. https://doi.org/10.1038/s41586-018-0561-9

Supplementary information accompanies this Comment. https://doi.org/10.1038/s41586-018-0561-9

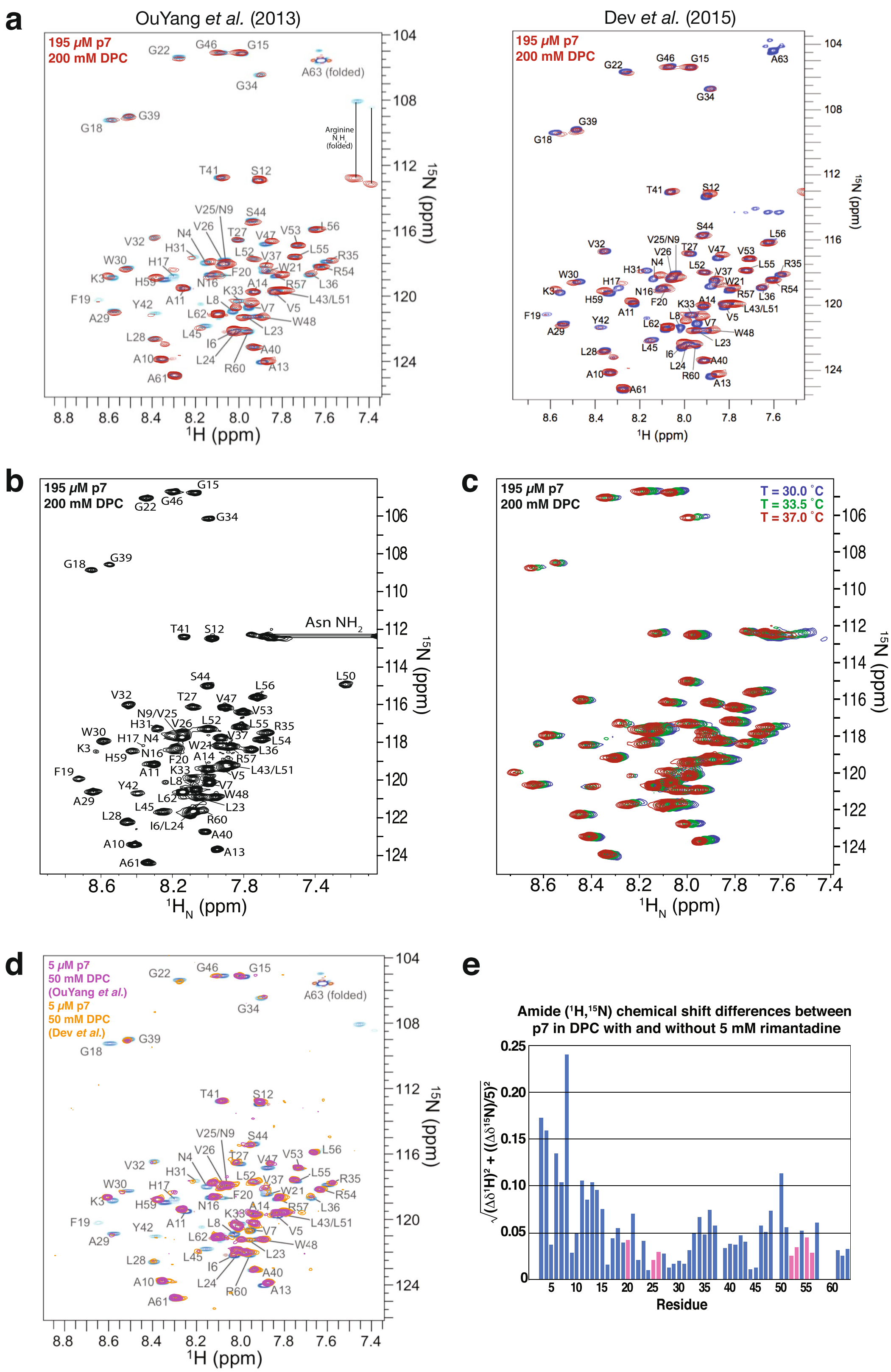
Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to N.Z. or J.R.S.

https://doi.org/10.1038/s41586-018-0561-9

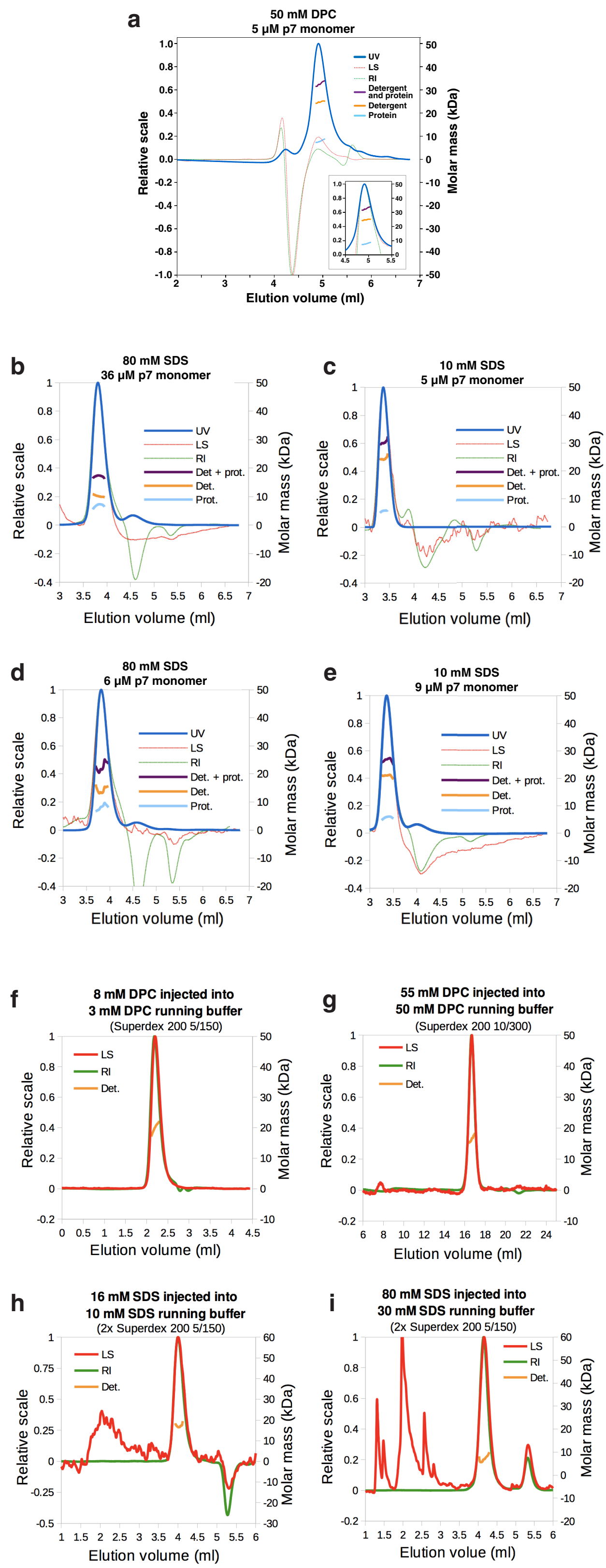


Extended Data Fig. 1 | Molecular dynamics of p7. a, **b**, Insertion of the proposed hexameric p7 structure[2] into lipid bilayers. MemProtMD[6] prediction for the hexamer insertion into a hydrated 1,2-dipalmitoyl-*sn*-glycero-3-phosphatidylcholine (DPPC) bilayer (**a**). The p7 structure[2] after insertion into a hydrated 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphatidylcholine (POPC) bilayer and simulated for 60 ns (**b**). Severe deformations and thinning defects of the bilayer can be seen, resulting in a large number of water molecules within the hydrophobic region of the bilayer. Water is shown as van der Waals spheres for oxygen (red) and hydrogen (white). For DPPC and POPC lipids, phosphorus and choline nitrogen positions are indicated with orange spheres and green spheres, respectively. In **b**, p7 α-helices and 3_{10}-helices are shown in magenta and blue, respectively. **c**–**e**, Simulations of monomeric p7 in 300 mM DPC at a protein to detergent ratio of 1:250. Independent 100-ns simulations of the horseshoe-like subunit conformation of the proposed hexameric p7 structure (**c**, **d**). At the end of the simulation, ~170 (**c**) and ~120 DPC (**d**) molecules, were observed bound to the protein. A hairpin conformation of p7 was simulated for 100 ns, at the end of which ~100 DPC molecules were observed bound to the protein (**e**). In **c**–**e**, p7 α-helices and 3_{10}-helices are shown in magenta and blue, respectively; the geometric centre of the DPC headgroup is indicated by an orange sphere, and the DPC hydrocarbon chain as yellow sticks. Only those DPC molecules bound to p7 are shown. The simulations in **b**–**e** were performed with the CHARMM36 all-atom force field using the protocol described[19].



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | NMR spectroscopy of p7 in DPC. a, Comparison of NMR spectra of p7, prepared as described[2], with previously published spectra. Left, 2D ^{1}H–^{15}N-TROSY spectrum of a sample of p7 in DPC produced in the present study (red cross peaks), recorded at 30 °C with a ^{1}H frequency of 600 MHz, overlaid with the spectrum published by OuYang et al.[2] (blue cross peaks). Right, our p7 spectrum (red cross peaks) overlaid with another published spectrum[8] (blue cross-peaks) to illustrate sample-to-sample variation. In both cases, the spectra were overlaid manually. **b**, ^{1}H–^{15}N HSQC of p7 in DPC at 37 °C. The spectrum was recorded at 600 MHz (^{1}H) on a Bruker spectrometer equipped with a CryoProbe. **c**, Temperature sensitivity of the spectrum of p7 in DPC. Spectra were recorded at three temperatures: 30 °C (blue), 33.5 °C (green) and 37 °C (red). The spectra have not been corrected for the temperature dependence of the ^{2}H-lock frequency. **d**, NMR spectra of p7 under SEC–MALS conditions. Spectra were recorded of 5 μM p7 with 50 mM DPC (detergent-to-protein ratio of 10,000) to match conditions at the SEC–MALS peak summit for a sample prepared as described[2] (violet) (see Fig. 1a and Extended Data Table 1), or as described by Dev et al.[8], with pre-gel filtration in 3 mM DPC buffer containing 100 mM NaCl (orange) (see Extended Data Fig. 3a). Before recording the NMR data, the sample in 3 mM DPC + 100 mM NaCl was dialysed against buffer (25 mM MES at pH 6.5, 3 mM DPC, and 1 mM DSS with 5% D_2O) overnight to remove salt, concentrated, and then diluted with NMR buffer (25 mM MES at pH 6.5, 50 mM DPC, and 1 mM DSS with 5% D_2O) to match the concentrations at the peak summit of the SEC–MALS elution (5 μM p7 and 50 mM DPC). Band-selective excitation short-transient transverse relaxation-optimized spectroscopy (BEST-TROSY) NMR experiments were recorded over ~20 h each on a 750-MHz spectrometer equipped with a CryoProbe to obtain sufficient signal-to-noise ratios for the dilute samples and overlaid with the published TROSY spectrum (supplementary figure 2a in OuYang et al.[2]). **e**, Backbone amide chemical-shift differences between a sample of p7 in DPC without and with 5 mM rimantadine. Data analysis was carried out on the spectra of OuYang et al.[2]. Chemical-shift differences were calculated as indicated on the vertical axis for the backbone amide resonances in the ^{15}N–^{2}H-labelled mixed-label sample and the ^{15}N–^{2}H-labelled sample containing 5 mM rimantadine. Residues in pink were previously identified[2] as forming the rimantadine-binding pocket (see figure 3c in ref. [2]). All reported concentrations are for monomeric protein.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | SEC–MALS of p7 and detergent alone. a, P7 in 50 mM DPC prepared as described by Dev. et al.[8]. The sample was run over two Superdex 200 5/150 columns in series in order to resolve the protein and micelle complex. The calculated masses were 7.8 ± 1.1 and 24.7 ± 0.6 kDa for protein and the associated detergent, respectively. The DPC concentration and maximum p7 monomeric concentration eluted (summit of peak) are indicated above the graph. **b**–**e**, SEC–MALS of p7 in 80 mM and 10 mM SDS. The SDS concentration and the maximum concentration of eluted monomeric p7 (summit of peak) are indicated above each graph. p7 is monomeric in SDS, as shown by its mobility on SDS–PAGE[2] (and our data, not shown). SEC–MALS molar mass analysis indicates, unambiguously, and in all cases, a monomeric state for p7 in SDS. The analysis also confirms the trend seen with DPC (Fig. 1 and **a** above), in that the amount of detergent associated with the protein is markedly and consistently higher for the samples at the lower detergent concentration (10 mM SDS) compared with samples in higher detergent concentration (80 mM SDS) (see data summary in Extended Data Table 1). Samples were run through two Superdex 200 5/150 columns connected in series for increased resolution. Sample injection volumes were 30 μl. Running buffer contained the indicated SDS concentrations and 10 mM sodium phosphate at pH 7.2. **f**–**i**, SEC–MALS of DPC and SDS micelles in the absence of protein. The detergent, its concentration in the injected sample and running buffers, and the chromatography columns used are indicated above each graph. As in **b**–**e**, SDS samples were run over two Superdex 200 5/150 in series to increase resolution. The $A_{280\,\mathrm{nm}}$ in samples without protein is negligible. Running buffers are the same as those used in experiments with protein (Fig. 1 and **a**–**e**), and the injected samples have higher detergent concentration to enable detection above the baseline. In **i**, the running buffer contained 30 mM SDS because the smaller size of micelles in 80 mM SDS in the absence of protein resulted in low signal-to-noise scattering. Molar masses (Det, orange line, right axis) increase slightly (by ~3 kDa for DPC and ~10 kDa for SDS) at concentrations close to the critical micelle concentration. Right axes: molar masses calculated at each point for protein (Prot; if protein is present), associated detergent (Det), and the detergent and protein complex (Det + Prot; if protein is present). Left axes: UV, $A_{280\,\mathrm{nm}}$; LS, light scattering; RI, refractive index. Detector signals are scaled to enable comparison. In **a**, the inset shows the detector signals normalized at the p7 peak. Extended Data Tables 1 and 2 show a summary of experimental conditions and mass calculations. The reported masses denote the value at the peak summit and the error is taken as the maximum difference from this value across the elution volume for which the molar mass is plotted. For samples with p7, negative and positive scattering and refractive-index peaks following the protein peaks are the result of distortions of the baseline upon sample injection that causes disequilibrium of detergent micelles in the running buffer[20].

Extended Data Table 1 | Summary of conditions and results for SEC–MALS studies of p7 in DPC and SDS

		Pre-injection			Eluted (peak summit)		Analysis of eluted p7 peak	
	Running buffer (mM)	p7 monomeric concentration (μM)	Det./Prot. (mol/mol)	Sample injection (μl)	p7 monomeric concentration (μM)	Det./Prot. (mol/mol)	Protein (kDa)	Detergent bound to p7 (kDa)
Fig. 1a	50 mM DPC	200	250	70	5	10000	6.4 ± 1.1	23.5 ± 0.5
Fig. 1b	3 mM DPC	15	200	15	0.9	3333	8.4 ± 0.7	42.4 ± 3.7
Ext. Data Fig. 3a	50 mM DPC	140	357	50	5	10000	7.8 ± 1.1	24.7 ± 0.6
Ext. Data Fig. 3b	80 mM SDS	464	172	30	36	2222	7.3 ± 1.8	10.2 ± 0.7
Ext. Data Fig. 3c	10 mM SDS	58	172	30	5	2000	5.9 ± 0.8	24.1 ± 2.3
Ext. Data Fig. 3d	80 mM SDS	58	1379	30	6	13333	8.3 ± 1.7	13.2 ± 3.1
Ext. Data Fig. 3e	10 mM SDS	120	83	30	9	1111	5.9 ± 1.2	21.0 ± 1.5

Data from Extended Data Fig. 3b, c show that at similar detergent/protein molar ratios the molar mass of detergent associated with p7 is approximately double in 10 mM SDS compared to in 80 mM SDS. Data from Extended Data Fig. 3d, e show that the result is consistent when changing the protein concentration. Together with the data for p7 in DPC (Fig. 1), the results suggest that at detergent concentrations close to the CMC (~8 mM for SDS and ~1.5 mM for DPC), a larger number of detergent molecules associate with p7. Running buffers contained the indicated detergent concentrations and 25 mM MES at pH 6.5 (DPC samples) or 10 mM sodium phosphate at pH 7.2 (SDS samples). The chromatography columns were Superdex 200 of sizes 10/300 (Fig. 1a) and 5/150 (Fig. 1b) or two 5/150 columns attached in series (Extended Data Fig. 3a–e), depending on sample volumes and concentrations, and resolution requirements. The samples (pre-injection) contained the same detergent concentration as the running buffer. Reported masses denote the value at the peak summit and the error is taken as the maximum difference from this value across the elution volume for which the molar mass is plotted in the SEC–MALS figures.

Extended Data Table 2 | Summary of conditions and results for SEC–MALS studies of DPC and SDS detergents in the absence of protein

Sample			Respective detergent concentration in running buffer (mM)	Micelle molar mass (kDa)	Superdex 200 dimensions
Detergent	Concentration (mM)	Volume (µl)			
DPC	8	5	3	20.2 ± 3.2	5/150
DPC	55	100	50	16.8 ± 1.8	10/300
SDS	16	30	10	17.2 ± 3.4	2x 5/150
SDS	80	30	30	6.9 ± 3.1	2x 5/150

Data are from experiments shown in Extended Data Fig. 3f–i. The micellar molar masses calculated from the SEC–MALS data at concentrations close to the CMC (~8 mM for SDS and ~1.5 mM for DPC) are in close agreement with those reported in the literature: ~19.0 kDa for DPC (Anatrace measurement in collaboration with M. Foster, University of Akron) and ~17.3 kDa for SDS[18]. At higher detergent concentrations, the micelle molar masses decrease by ~3 kDa for DPC and ~10 kDa for SDS. Running buffers contained the indicated detergent concentrations and 25 mM MES at pH 6.5 (DPC samples) or 10 mM sodium phosphate at pH 7.2 (SDS samples). Reported masses denote the value at the peak summit and the error is taken as the maximum difference from this value across the elution volume for which the molar mass is shown in the SEC–MALS figures.

Chen et al. reply

REPLYING TO B. P. Oestringer et al. *Nature* **562**, https://doi.org/10.1038/s41586-018-0561-9 (2018)

In the accompanying Comment[1], Oestringer et al. argue that the p7 protein from hepatitis C virus, which we used to determine the p7 oligomeric structure in dodecylphosphocholine (DPC) micelles[2], is monomeric. Here we show, with direct experimental evidence, that their assertion is wrong.

In our previous study[2], the hexameric state of p7 in DPC was indicated by two pieces of evidence. First, the negative-stain electron-microscopy analysis of the sample generated 2D reference-free class averages that are clearly hexameric. Second, the sample prepared in the same way, containing 1:1 mixture of ^{15}N–^{2}H-labelled monomer and ^{13}C-labelled monomer, showed unambiguous nuclear Overhauser effects (NOEs) between the amide protons of the deuterated monomers and the methyl protons of the fully-protonated monomers, indicating that the p7 reconstituted in DPC formed oligomers.

Oestringer et al.[1] argue that the ^{15}N–^{2}H-labelled p7 in the mixed sample was incompletely deuterated, and therefore the inter-monomer NOEs that we reported were actually intra-residue NOEs (figure 2 in ref. [1]). In our study, we recorded three ^{15}N-edited NOE spectroscopy (NOESY) spectra: one with a mixed sample containing 1:1 mixture of ^{15}N–^{2}H-labelled p7 and ^{13}C-labelled p7, and the other two with samples containing only ^{15}N–^{2}H-labelled p7 but in the presence of the drug amantadine or rimantadine. The latter two spectra are from samples that are not isotopically mixed, and can therefore be used as negative controls. The key question is whether or not the six most prominent inter-monomer NOEs reported (supplementary figure 8 in ref. [2]) are also present in the two negative controls. These NOEs relate to residues 7, 10, 11, 14, 23 and 24, but none of these are shown in figure 2b of Oestringer et al.[1]. We show that none of the six inter-monomer NOEs in the mixed sample are present in the negative control samples (Fig. 1), indicating that they have arisen from oligomerization of different isotopically labelled p7 monomers. We note that an unambiguous, complementary experiment to the mixed NOE experiment is measurement of mixed paramagnetic relaxation enhancement (PRE) using a sample in which half of the monomers are spin-labelled at the N terminus and the other half are ^{15}N-labelled. If the p7 in DPC forms oligomers, the NMR resonances of the relevant region should show strong PRE. Oestringer et al.[1] did not perform this simple and direct measurement to prove that the p7 is monomeric in DPC.

Oestringer et al.[1] presented data from NMR relaxation and size-exclusion chromatography coupled to multi-angle light scattering (SEC–MALS) studies to support their conclusion. NMR-derived rotational correlation time (T_c) is an ambiguous indicator of protein size owing to the presence of mobile regions. If T_c is to be used to infer the size of the p7 complex, then the mobile regions such as the first helix (H1) and the C-terminal residues (57–63) should be excluded. It has been shown that the average T_c at 30 °C excluding these regions is ~22 ns (supplementary figure 2 in ref. [3]), much larger than the 10.1 ns reported[1] by Oestringer et al. A T_c of 22 ns would be consistent with a 60–70 kDa complex. SEC–MALS analysis of small membrane peptides in detergent micelles is convoluted and unreliable, reflected by the inconsistency observed by Oestringer et al.[1] that the micelle sizes at different detergent concentrations are very different.

Oestringer et al.[1] also used other NMR-based arguments that are much less convincing. They argue that the molecular alignment tensor derived from residual dipolar coupling (RDC) measurements should be axially symmetric if p7 is a hexamer. However, this is only true if the p7 hexamer is completely rigid and symmetric. NMR relaxation data already indicates that large internal dynamics exist[3] and there is no a priori reason to suppose that these motions are symmetrical.

In conclusion, Oestringer et al.[1] present multiple indirect arguments that p7 is monomeric in DPC, but fail to convincingly refute the obvious inter-monomer NOEs, which represent direct evidence of oligomeric assembly. Their conclusion is further challenged by a recent study showing that the oligomeric assembly of p7 in DPC is consistent with that in bicelles that closely mimic the membrane environment[3].

Wen Chen, Bo OuYang and James J. Chou are solely responsible for this Reply; other authors from the original Letter who were involved in structural and functional studies of p7 did not contribute to this response and are not listed here. Readers are welcome to contact the authors for further unpublished data including intermolecular NOE and PRE data.

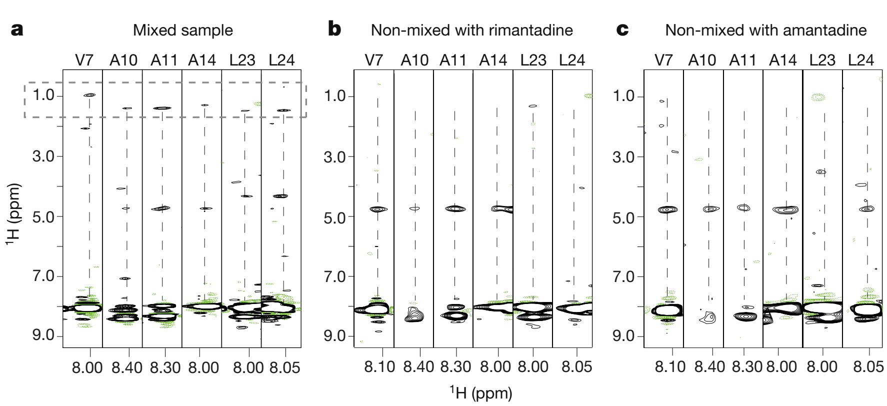


Fig. 1 | Mixed NOEs and negative controls. **a**, NOE strips from the mixed sample containing 50% ^{15}N–^{2}H-labelled p7 (5a) and 50% ^{13}C-labelled p7(5a). Inter-protomer NOEs are indicated in the dashed box. **b**, **c**, NOE strips from non-mixed samples containing 100% ^{15}N–^{2}H-labelled p7(5a) and the drug rimantadine and amantadine, respectively.

BRIEF COMMUNICATIONS ARISING

Wen Chen[1], Bo OuYang[2] & James J. Chou[1]*
[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA. [2]State Key Laboratory of Molecular Biology, National Center for Protein Science Shanghai, Shanghai Science Research Center, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai, China. *e-mail: james_chou@hms.harvard.edu

Published online 17 October 2018.

1. Oestringer, B. P. et al. Re-evaluating the p7 viroporin structure. *Nature* **562**, http://doi.org/10.1038/s41586-018-0561-9 (2018).
2. OuYang, B. et al. Unusual architecture of the p7 channel from hepatitis C virus. *Nature* **498**, 521–525 (2013).
3. Chen, W. et al. The unusual transmembrane partition of the hexameric channel of the hepatitis C virus. *Structure* **26**, 627–634 (2018).

Author contributions W.C., B.O. and J.J.C. wrote the response.

Competing interests Declared none.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.J.C.

https://doi.org/10.1038/s41586-018-0562-8

CAREERS

CONFERENCE QUESTIONS Men ask them, women don't. Why? p.451

EDUCATION China's appeal explained p.451

HARASSMENT TALE Researcher addresses immigration rally p.451

ADAPTED FROM SORBETTO/GETTY

BEHAVIOUR

Tackling harassment

Three real-life stories of online abuse — and how scientists got through it.

Researchers who study topics such as climate change and vaccines can become targets of online behaviour ranging from threatening e-mails to coordinated social-media attacks. *Nature* asked researchers who have been digitally harassed what they've learnt from the experience.

DAVID KEITH
Engage judiciously

Environmental scientist, Harvard University, Cambridge, Massachusetts

I do solar geoengineering experiments — notably, researching the chemical impacts of reflective particles that may be sprayed into the stratosphere to minimize incoming solar radiation interact with themselves and other compounds in the atmosphere. I make a distinction between harassers — people who send me more than 100 e-mails per year — and people in the mainstream environmental-science community who don't agree with my research.

I don't always engage with harassers. I mostly ignore the harassing tweets. The e-mails are harder to ignore — they seem more personal, so I do respond to quite a few, and sometimes I can change the senders' minds.

Over the past decade or so, I've been harassed by people who believe in the 'chemtrail' conspiracy theory — which proposes that long-lasting condensation trails left behind by aircraft are evidence that governments deliberately spray chemicals for nefarious purposes. Around 20–30% of the US population takes seriously the idea that these purported chemical releases might be for solar-radiation management, human-population control or chemical warfare. I estimate that about half of all tweets around solar geoengineering are in connection with chemtrails.

Routinely, I receive violent, sometime hideously anti-Semitic voicemails, e-mails and letters. A decade ago, I called campus security twice when the harasser became threatening, but nobody has ever been physically violent.

There is a huge gap between online rage and in-person rage. That said, security people at my institution routinely install office alarms, and they advised me to take common-sense steps — for example, to lock my door and pay attention to strangers. I also have conversations with conference organizers before meetings to make sure somebody knows the phone number for campus police in case there is a threat.

When someone sends a hateful thing, I'll ask if that made them feel good. I also ask why they think I'm evil and that I "murder kids". I remind them that I'm a human being, and that I have kids, too. I tell them that I think they've been fooled by some nonsense on ▶

the Internet, and that they are welcome to talk to me about climate change or geoengineering experiments designed to mitigate climate change. A couple of times, the aggressor has apologized.

The biggest challenge for democracy is learning how to lessen the number of people who believe things that are objectively wrong. I don't think that hiding from it and pretending it isn't there is a good idea. Vilifying people who hold those ideas is not a good approach either. We can't see them all as the enemy.

Instead of trying to change people's minds by telling them they are wrong because an expert says so, I try to question them in a way that shows I take their concerns seriously and reveals how their argument falls apart under scrutiny. In the case of chemtrails, I ask where the supply chains are for the poison, how the dispersal devices are engineered, how all of this has been kept secret for so long and, finally, what the motive is.

Approach debates with caution. They can be useful — but scientists are accustomed to ground rules of honesty and logic in debates, and it's tough to debate with people who are not using honesty and logic. Don't panic if you're being harassed online. The harassment ultimately is not about you, even if it seems personal. Be judiciously willing to respond.

Still, in the end you might have to make a decision. I have upfront conversations about this with postdocs and graduate students, and I encourage them to think through the pros and cons of working in this field. The upside is that it's a new, growing field. The downside is the criticism and polarization. I warn people that hard policy debates are part of this field right now, and that if students don't want to be involved, it might not be the right field for them.

JOANNA HAIGH
Play it straight

Atmospheric physicist, Imperial College London

Harassment, usually by e-mail or attacks through blogposts, comes in waves. I probably get about 100 messages a year. It usually follows statements I've made on the radio or in the press about climate change, or after something has appeared on a climate-change denier website. It can be a range of things — from "You've got it all wrong" and "You are making all of this up" — to extremely rude, offensive personal attacks.

Many of the comments about me have gendered overtones, referring to me as "priggish" or "that woman", or telling me to "stick with flower-arranging". The people who give their names — and many don't — are always men. The worst offender doesn't give a name and has sent about a dozen multi-page screeds.

I have rules of engagement. I try to engage — but only with people who haven't been offensive. I have a brief fact sheet on the truth about global warming. If they ask scientific questions, I take a stab at answering them. I never respond to anything personal. I have had one or two write back and thank me for clarifying. Responding to these messages takes a lot of time and energy. At times, it can be a whole day's worth of answering.

Because of the time it takes and the harassment, I am not on Twitter. I know people who do a great job on Twitter, and I'm pleased they take it on. I don't think we can ignore people without being labelled arrogant. I am paid by the public purse, and I have a responsibility to explain to people about the work I do.

I worry about younger scientists who can find themselves targets for attacks they are unprepared to handle. My advice is simple: play it straight. Don't rise to the bait. Explain politely what you understand and what perhaps they have misunderstood. If they are offensive, do not respond.

CHRISTINE LATTIN
Be transparent

Environmental physiologist, Louisiana State University, Baton Rouge

In 2017, while I was a postdoc at Yale University in New Haven, Connecticut, I started getting e-mails that claimed that my research was cruel and pointless. I use wild birds to study stress hormones and neurotransmitters. An organization made misleading claims about my work that led to hundreds of harassing messages. Some included death threats.

It has been stressful and challenging, but these harassers' efforts to shut down my research and to silence me have not been successful.

If offensive, do not respond, advises Joanna Haigh.

IMPERIAL COLLEGE LONDON

I was advised to let it blow over and not respond, but that didn't seem to make things better, and it might have made them worse. I decided to defend myself and get different information out there regarding these claims. So I started talking to journalists about my work and speaking up on social media. Taking ownership of my own story made me feel like less of a victim. It's crucial to be open and transparent about our work and advocate for its importance.

I address the false claims directly when possible. I make clear how and why I do this work and that those of us doing animal research receive a ton of oversight. I explain that a lot of people are in place to make sure the animals are taken care of, that suffering is minimized and that the research is justified. For example, every study I do is approved by a university Institutional Animal Care and Use Committee, and both universities I have worked for are accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International. All my research complies with the Ornithological Council's Guidelines for the Use of Wild Birds in Research.

The worst harassment I've had was on Facebook, so I unplug from social media and spend time with my family, friends and pets.

I also reassessed my professional web page. Although I thought I was being so open by making my papers available and creating a statement of my research, the language on my website was technical and not accessible to people at all. I got rid of the jargon and worked with a communications professional to explain clearly the reasons for my research.

I also have a 'frequently asked questions' section to address specific, often-repeated claims, such as 'animal research is unnecessary'. In my response, I point out that although non-animal methods such as cell culture or computer models can be excellent, they have limitations. I also share how I have pioneered less-invasive ways of studying stress as well as new imaging techniques for studying the brain and body. That is the most visited portion of my website. Now, if people Google me, they see two sides of the story.

Do not be afraid to ask for or accept help. I study stress. Exercise helps you to cope with stress. Tell people about what is happening to you and get support from family, friends, colleagues and current and former principal investigators. I have received a lot of messages of support, which has really helped.

There are also specific organizations — Speaking of Research, for example — that can offer support. That group helped me to put together rebuttals to the campaign organization's claims. And its director reminded me not to take the harassment personally, because it isn't about me. ■

INTERVIEWS BY VIRGINIA GEWIN

Interviews have been edited for length and clarity.

TURNING POINT
Immigrant defender


WISTAR INSTITUTE

Born in Sri Lanka, Ashani Weeraratna was raised in Lesotho in southern Africa and moved to the United States in 1988 to pursue an undergraduate science degree. Now a skin-cancer researcher at the Wistar Institute in Philadelphia, Pennsylvania, she has experienced harassment during her three decades in the country. An escalation in incidents prompted her to address a rally protesting against family separations.

Describe your most recent experience of harassment.
In April, I was in a grocery checkout line when someone told me that people like me are from shithole countries and live like animals. In January, I posted on Twitter about being a principal investigator and mentor, and someone asked why I was not taking my science back to my home country. I explained that I am a citizen, that I think the United States is the best place to do science, and that my husband and daughter also live here. The person told me I should leave jobs available for US postdocs. I blocked her when she took a screenshot of my profile photo and one of me with my daughter, whom she referred to as an 'anchor baby' (a pejorative term used to describe a child born in a country with birth-right citizenship to a non-citizen mother).

Was the harassment different before the change in US administration in January 2017?
As a more-junior scientist, I thought of myself as an overlooked voice — a woman of colour doing science. It wasn't so much harassment before then as it was not being taken seriously. There were instances, for example after the terrorist attacks on 11 September 2001, when a couple of random people spat at me and said horrible things like, "Go home, dirty Arab." I tried not to let it bother me back then. It's different now, though, because I have a biracial daughter to protect.

How did you come to speak at a pro-immigration rally?
The Trump administration's policies — including the ban on travel from some Muslim-majority countries and possibly rescinding the visas that allow spouses of immigrants to work — are affecting science. Border policies that separate families and incarcerate children, as well as the dehumanizing, divisive language, also bother me. I had one talented collaborator from Syria who was at Wistar for four years on a work visa. But now her visa keeps getting denied. When our Democratic state senator, Daylin Leach, asked me to speak at a rally on 30 June to support immigrants and protest against family separations, I had to think about it. But I decided the public needed to understand that these policies hugely affect biomedical research. More than 40% of the US cancer-research workforce is made up of immigrants. At Wistar, 289 employees are from more than 20 different countries.

Were you concerned about participating?
I worried that, by speaking out, I could jeopardize my federal grants, which support my lab, my students and my institute. I received legal advice that as long as I spoke as a private citizen, I'd be fine. I also talked it over with my husband, a cautious person, who said I needed to be on the right side of history.

What was the reaction to your speech?
The positive response was overwhelming. I posted a video of my talk on Facebook and friends encouraged me to make it public. It's had more than 4,000 views so far.

What did you share at the rally?
I grew up in landlocked Lesotho. To get medical training, I would have had to go to South Africa, a country that had apartheid and was segregated in the 1980s. I saw limited opportunities to pursue my dream of being a cancer researcher as a woman of colour there. When I came to the United States for college, to my mind, the country was a bastion of free speech and a great melting pot. To feel like that's being reversed so quickly is frightening and discouraging. I implored politicians to do what they can to ensure that the American dream doesn't become an American nightmare. ■

INTERVIEW BY VIRGINIA GEWIN

This interview has been edited for length and clarity.

EDUCATION

China calling

Universities and research institutions in China that have reputations of excellence, or are highly ranked by external organizations, are among the draws for undergraduate and postgraduate international students who are flocking to the country, according to a report in the *Journal of Studies in International Education* (W. Wen & D. Hu *J. Stud. Int. Educ.* http://doi.org/cvfs; 2018). The study finds that since 1995, the number of international students in China has grown 12-fold, from 36,855 to 442,773. More than half of those students are from other Asian nations. The authors found that these students, apart from those from Japan, were more concerned with the reputation or ranking of Chinese institutions than were their counterparts from North America, Europe and sub-Saharan Africa. The study, based on survey results and interviews with 30 international students, also found that China's cost of living, admission policies for international students and scholarship programmes increased the nation's appeal to foreign students. The government offers roughly US$300 million in scholarships to international students each year.

CONFERENCES

Silence is not golden

That male speakers outnumber female speakers at seminars and conferences has been a long-standing issue in science, but a gender gap exists on the other side of the lectern, too. Male conference attendees are about 2.5 times more likely than their female counterparts to ask questions of a speaker or panel after a presentation, a study in *PLoS One* has found (A. Carter *et al. PLoS One* **13**, e0202743; 2018). The authors collected observational data at 247 departmental seminars, hosted at 35 institutions in 10 countries. They also carried out an online survey to gauge how researchers felt about asking questions. By analysing the responses from 509 researchers around the world, the authors found that women were significantly more likely than men to say that they had kept silent because they were unsure whether their question was appropriate, or because they did not have enough "nerve" to ask it. Lead author Alecia Carter, a behavioural ecologist at the University of Montpellier in France, and her co-authors suggest that women might be more likely to raise their hands if organizers allotted more time for questions, or scheduled a short break first for attendees to gather their thoughts.

FERROMAGNETISM

Strange attraction.

BY D. A. XIAOLIN SPIRES

"I could never understand your return to nativism," said Gramps. He sprayed his joints with silicone fibre and waited as they stiffened into rubbery cords. In an instant, the cords vanished.

"Gramps, you know not to use that stuff in front of me," I said, feigning a cough I hoped was convincing. The smell of strange ether filled the workshop as I dressed the heavy-elemental skin on the dense skull before me.

Gramps waved the spray bottle in the air, almost toppling my prized tin of Atomic Number *^#~ powder.

"Watch your hands," I said, throwing him a warning glance. "I had to emerge at the Castaway Cluster to get that. It'd take me months to earn enough tickets to disembark in that region again."

"Intrauniversal travel," he snorted. "A waste of time. All the amazing things lie beyond." Gramps now held the spray to his spindly, metal legs. A tingle excited my face. I wheezed.

"Allergies to extrauniversal matter don't exist. It's a scientific fact," he said. "Your body doesn't have any negative interactions with it." He gave his feet a few spritzes.

"Maybe in your generation. Who knows what discoveries are still to come. All I know is that the stuff is toxic," I croaked out a few spasmodic air expulsions for good measure. Gramps shook his head, but pushed the button that spread a sealant over the spray nozzle.

"There, happy now?" he said, setting it down.

I nodded. For a moment, quiet filled the workshop as I lowered the laser down over the disembodied head. A groan escaped the specimen as I affixed the rest of the body that I had dragged over. Gramps turned in my direction and watched.

I latched on the head with clamps. Something felt tight as it caught. A rush of the parts, head and neck met each other, as if in longing.

The bot, yet another copy of me — young and ductile rather than the creaky copy Gramps consisted of — exposed its sensory orifice to the world as it lifted its alloy covering. Its sensory cavity was round and luminous, with a smooth bead within, turning about in wide scans of the room.

I felt a tugging sensation from deep within pulling me towards it. I studied its face, scanning in detail its vulnerable sensory bead, with a tenderness I've never experienced before. I could feel myself approaching its body, putting my limb on one of its own, one metallic arm to another, an expression of homologous intimacy. I've never felt so connected to any of my progeny before.

➲ NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

Its sensory bead shone with a kind of awareness nonexistent in newborns.

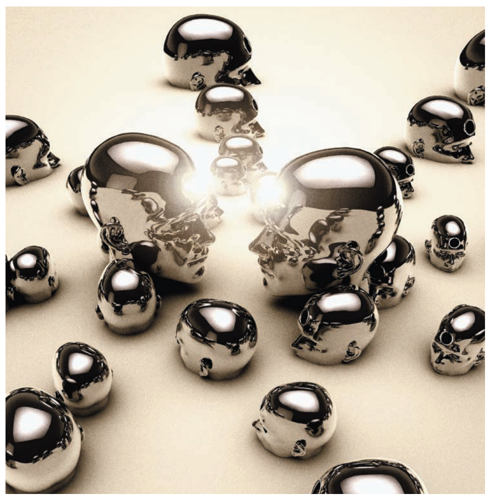


ILLUSTRATION BY JACEY

Gramps whistled.

"What did you do to it?" I whispered at Gramps. He looked away. An unreadable expression.

He turned back to the newborn, as if unable to help himself.

"Oh, nothing," he said. His pursed face suggested otherwise. He blinked aside his alloy covering, exposing a dilated sensory bead of his own.

"I'll take it apart —"

"You wouldn't," said Gramps.

"You want to test me?" I started to unhinge a lower joint. A pain shot through my own.

"I just inserted a variable that instilled uncertainty."

"A variable?"

"Yes, like I did to you."

My hands stood still as I grasped its lower joint. "What're you talking about?"

"You, your impassioned will and fire. Your grit and persistence. All from just a bit of magnetism."

"Magnetism? Wait, *magnetism*? That requires, what is that substance called? *Iron*. But that's —"

"Yes, extrauniversal material."

"— illicit. Are you kidding? It's in me?"

Gramp's sensory bead transmitted a deliberate sheepish expression.

"If so, that makes me a —"

"You're a mutt alright," he proclaimed, in a voice too loud. He was proud of himself.

I could feel my joints pull towards my workmanship, my exact duplicate, with an uncanny sensitivity I had never encountered towards my creations before.

"I put in an opposing magnetic force in this one. You'll feel what the other universe calls 'parental affection.'"

I looked into this copy, at its sensory bead that mirrored my own. No wonder Gramps went crazy after my father died. I saw now that he must've implanted something like this into himself. I felt a sickness pass through me as I realized my very being violated the nativist dictum. A mutt. Its very notion made me want to disgorge liquid alloy.

Gramps shifted his leg, grabbed the silicone spray and spritzed its strange twining tendons onto both his bottom limbs. He was allowed to use that stuff; he was grandfathered in for use of restricted materials.

I felt the wrenching feeling that drew me to my newly born copy. I grabbed the bottle. I shook it. It rattled. I spritzed the newborn's legs, chest, face and even its sensor with a generous dose. The spray had no properties from Atomic Number * to Atomic Number *##@%~. It was a wrongful substance. The silicone twisted into fibres before vanishing into thin air, an invisible net whose purpose seemed to offer structural support.

The copy closed its eyes and — was I imagining it? Or did it let out an approving sigh?

"The magnets interact with the silicone, making it hard to trace," said Gramps. He patted his great-grandson's head. "No one will be any the wiser."

"Except me," I said. I sprayed my own face, wondering how many times Gramps must've applied similar spray to me without my knowing. "A bastard child of a bastard son."

"The best there was," said Gramps, his sensory bead registering faraway longing. He got up with what looked like renewed vigour and tucked the spray bottle away into a dark recess of the otherwise brightly lit workroom. ■

D. A. Xiaolin Spires *spritzes Atomic Number *^#~ behind her ears before stepping out. Work in* Clarkesworld, Analog, Fireside, Grievous Angel, Galaxy's Edge, LONTAR, Terraform, Andromeda Spaceways *and various anthologies. daxiaolinspires.wordpress.com Twitter: @spireswriter.*

From the ground up ... and up

Industry and investment are powering WESTERN AUSTRALIA'S RESEARCH in terrestrial, marine and cosmic science

Long recognized as an attractive research environment, Western Australia boasts ancient geological formations, a World Heritage-listed barrier reef, a rich collection of flora and fauna, and perfect conditions for radio astronomy. It is a setting that fosters a sense of wonder, a mindset for discovery and a desire to innovate.

Research diversity

Western Australia's research excellence spans a wide array of disciplines, from radio astronomy and supercomputing to marine science and biodiversity; from minerals and energy to health translation. Central to this research ecosystem are the state's five universities: Curtin University, Edith Cowan University, Murdoch University, The University of Notre Dame and The University of Western Australia.

Across the five universities, AUD$984 million was invested in research and development in 2016. Their projects span from the subatomic scale to the bounds of the known Universe, with outcomes geared towards solving industry problems or extending the limits of knowledge.

Reflecting the state's broader population, Western Australia's universities are culturally diverse environments that encourage participation and collaboration irrespective of culture, religion, gender and age. There are more than 4300 active researchers in the state's universities, with many collaborators in Western Australia's hospitals, independent medical research institutes, government organizations, and industry and community sectors. Western Australia has more than 6000 research students who will continue to fuel discovery and research for decades to come.

National support

These efforts are supported nationally by strong investment in research infrastructure. For example, the state capital, Perth, hosts the largest public supercomputer in the southern hemisphere, the Pawsey Supercomputing Centre, which is used for data-intensive research in astronomy, physics, geoscience, health and economics. It recently received AUD$70 million from the Australian government to secure the next generation of supercomputers, expected to become operational in 2019–2020. Another example is the new federally funded Cyber Security Cooperative Research Centre, which brings together the state's cyber security researchers and experts, and complements the Academic Centre of Cyber Security Excellence and the Joondalup Innovation Hub with a focus on cyber.

Strategic collaborations

The integrity and strength of Western Australian research is evidenced by numerous world-class partnerships. A significant amount of research undertaken in Western Australia is driven by industry needs in the resources sector, production industries, health and biomedicine, information and communications technology, space, astronomy and data science. Strategic collaborations bring interdisciplinary expertise to problems of considerable scale, complexity and duration. These partnerships enable research teams to be confident that project outcomes will translate into tangible benefits for industry and community partners.

Western Australia continues to invest in

Beneath southern skies: life, art and radio astronomy

AUSTRALIA'S IMPACT IN GALACTIC AND EXTRAGALACTIC ASTRONOMY has soared since 2009 when two Perth universities established an international research hub 'near' a radio-quiet desert region

In less than a decade, the International Centre for Radio Astronomy Research (ICRAR) has ballooned into a 190-strong cohort that is participating in the science, engineering and technologies of the world's most ambitious and anticipated radio telescope — the Square Kilometre Array (SKA).

ICRAR was established in 2009 by Curtin University and the University of Western Australia, with support from the Western Australian government. At the time, Australia and South Africa were developing competing bids to host the SKA. Being 'only' 800 kilometres from Australia's proposed site in the mid-west Murchison region, Perth was an ideal base for SKA-related research and development, particularly since it offered the combined science, engineering and data-intensive research capabilities of Western Australia's two largest universities.

Globally, more than 100 organizations in 20 countries are contributing expertise and support to the multi-billion-dollar SKA project. Once the first phase is operational, the full instrument will use over 100,000 low-frequency antennas in Western Australia and about 200 dish antennas in South Africa to map the night sky. This will enable astronomers and physicists to explore the early period of the Universe and learn more about gravitational waves, the formation of the first stars and galaxies, and other mysteries.

Using the new SKA-precursor telescopes at the Murchison Radio-astronomy Observatory and international sites, ICRAR researchers have discovered a lot about galactic and extragalactic phenomena, such as fast radio bursts and black holes. Their activities have contributed more than 1300 scientific papers in leading journals, including 19 publications in *Nature* and *Science*. An all-sky survey by the Murchison Widefield Array precursor created a new radio-source catalogue of 300,000 galaxies and a radio colour panorama of the Universe.

ICRAR has also facilitated a very different but equally vibrant and inspiring world first — a collection of astronomical artworks by Western Australian Indigenous artists. The artworks resulted from a campfire exchange of knowledge about the night sky between ICRAR researchers and Wajarri Yamatji people, the traditional custodians of the Murchison land. The Australian collection, titled Ilgarijiri — Things Belonging to the Sky, has joined similar artworks from South Africa to create the Shared Sky exhibition, which is touring the SKA member countries. ■

Advertiser retains sole responsibility for content

frontier science and large-scale collaborations that foster research excellence, engagement and impact. For example, the International Centre for Radio Astronomy Research was founded in 2009 to support Australia's bid to host the world's largest radio telescope and one of the largest scientific projects in history — the Square Kilometre Array.

Likewise, the Western Australian Health Translation Network is a focal point for health and medical research to capitalize on state government investment in health research, patient care and population wellbeing. Through the network, doctors and clinicians from four hospital foundations and more than 15 independent medical research institutes work with university researchers towards preventing, treating and curing medical conditions that afflict children, adults and the elderly.

The Australian National Phenome Centre is an international research node dedicated to developing and delivering metabolic phenotyping services to better understand the interactions between genetics, environments, microbiomes, diets and lifestyles across populations. The centre enables Australian scientists to participate in the affiliated International Phenome Centre Network, a global research programme that seeks to transform health and improve disease prevention, detection and treatment.

For many years, the Minerals Research Institute of Western Australia has promoted minerals research to support investment in and operation of a globally competitive minerals industry in Western Australia. The institute has a long history of successful collaboration with research and government bodies in Australia and overseas to promote minerals research.

The Magnus supercomputer at the Pawsey Supercomputer Centre

THE WORLD'S LARGEST RADIO TELESCOPE AND ONE OF THE LARGEST SCIENTIFIC PROJECTS IN HISTORY

The Western Australian Energy Research Alliance (WA:ERA) also brings targeted, industry-driven research and development to the fore. Through this alliance, university researchers work with Woodside Energy, Chevron Australia and Shell Australia to support future energy developments and emission reduction goals. WA:ERA's focus includes subsurface technologies and geosequestration-related research, supported by a vast suite of advanced characterization and analytical facilities and instruments in Western Australia's research institutions.

The Western Australian Biodiversity Science Institute serves the growing need for research that balances economic growth with best practice in environmental protection. Through the institute, four Western Australian universities and several government agencies contribute expertise in mine site rehabilitation, landscape conservation, urban diversity and areas relevant to sustainable development.

The Western Australian Marine Science Institution brings together 15 research, industry and government partners to create a leading Australian organization for research into marine ecosystems and the impact of aquaculture, resources and tourism industries. It has more than 250 scientists representing oceanography, biodiversity, aquaculture, biotechnology and climate science — an essential combination of expertise for ensuring the sustainability of Western Australia's magnificent coastline and marine ecosystems.

These research hubs and facilities represent a selection of Western Australia's research strengths. With the defined priority areas of mining, energy, health, food, environment, space, technology and STEM, Western Australia will continue to promote and support exciting opportunities for research collaborations and impact. ■

LOOKING EVER FORWARD

As Western Australia's largest and most internationally engaged university, CURTIN UNIVERSITY CONTINUES TO BUILD INFLUENTIAL RESEARCH COLLABORATIONS that strike a balance between demand-driven and researcher-driven research

Founded in 1966 during Australia's third mining boom, the Western Australian Institute of Technology embodied the spirit of industry collaboration and real-world endeavour. Within three decades, its impact on education and research saw it become a university. Named after Australia's 14th Prime Minister, Curtin University has embraced John Curtin's wisdom that a great university should look ever forward.

Today, Curtin University is Western Australia's largest and most culturally diverse university, ranked in the top 1% in the world by the Academic Ranking of World Universities. It has recently achieved top-100 status in several fields, including mineral and mining engineering, for which it ranked second in the QS World University Rankings in 2017 and 2018. The university has a global reach through its campuses in Malaysia, Singapore, Mauritius and Dubai.

Striking the right balance

Since its inception, Curtin University has aimed to strike a balance between research that meets industry demand and curiosity-driven research, an approach that is an integral part of the university's DNA. Critical to achieving this balance is the forging of strong partnerships between academia and industry. Under the umbrella of the Western Australia Energy Research Alliance, for example, Curtin University is collaborating with global companies (including Woodside Energy, Chevron and Shell Australia) to develop efficient exploration and recovery methods that minimize environmental impact. And its research strengths in theoretical physics, computer science and mathematics have led to valuable partnerships with NASA, Cisco, Optus and the Royal Australian Navy.

EMERGING TECHNOLOGIES ARE ESSENTIAL FOR DELIVERING RESEARCH OUTCOMES

In agriculture, recognizing the ever-increasing need to balance productivity with natural resource management, Curtin University has joined the Food Agility Cooperative Research Centre, a collaboration between 54 private businesses and universities that is applying spatial sciences and big data to drive digital innovation across farming and agriculture. The university also has the single largest partnership with the federally funded Grains Research and Development Corporation through its Centre for Crop and Disease Management, which researches fungicide resistance, molecular genetics and bioinformatics.

Exploring the skies

In recent years, Curtin University has partnered to drive rapid growth in planetary science and astronomy, the latter owing to its strengths in spatial sciences, theoretical physics, astrophysics and electrical engineering. The International Centre for Radio Astronomy Research (a joint venture between Curtin University, the University of Western Australia, and the Western Australian government) is paving the way for the world's most powerful radio telescope, the Square Kilometre Array, through the construction and operation of the Murchison Widefield Array and other precursor projects. These projects have proven significant in their own right, contributing to discoveries in galaxy evolution, black hole formation, and the accretion of the first structures in the Universe.

Curtin University's planetary science research team is also on the international radar, with their Desert Fireball Network — a network of autonomous observatories in remote Australia producing a vast catalogue of fireball trajectories, which will help researchers pinpoint the location of meteorites and understand their origin in the Solar System. A recent partnership with NASA's Solar System Exploration Research Virtual Institute has expanded the network's role to tracking objects other than meteorites, including the OSIRIS-REx spacecraft, as it streaked across the southern sky to acquire samples from the asteroid Bennu. Commercial applications are being developed in collaboration with Lockheed Martin for space situational awareness.

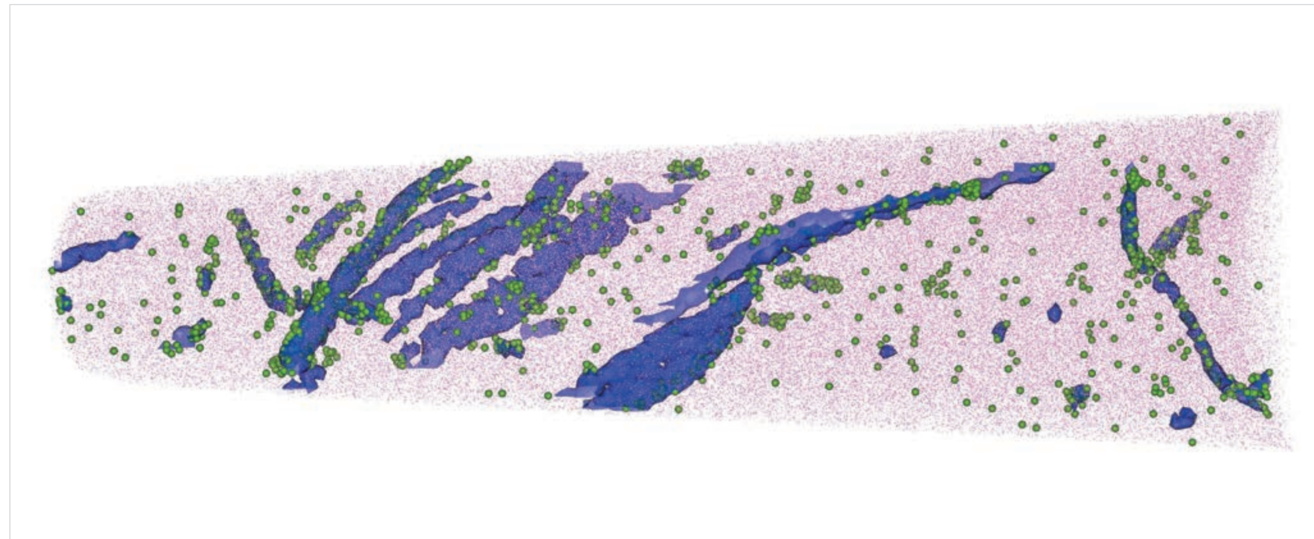
Employing emerging technologies

Emerging technologies are essential for delivering research outcomes. Curtin University is a world leader in geosciences, boasting one of the most advanced geological dating and characterization facilities in the world, the John de Laeter Centre. The Pawsey Supercomputing Centre supports the university's big-data projects, particularly those involving extensive simulation, data processing or

Curtin University has six campuses located around the rim of the Indian Ocean

Using research to enhance agriculture is a major focus at Curtin University

© LALS STOCK/Shutterstock



Curtin's Geoscience Atom Probe is helping Earth and planetary scientists trace the geological history of our planet and the solar system. This nanoscale rendering shows elemental segregation in a meteorite.

data storage, while the Curtin Institute for Computation focuses on data analytics and new visualization and modeling methods. At Innovation Central Perth, Curtin University researchers collaborate with experts from Cisco, Woodside, Data61 and several government departments and small enterprises to find solutions for cloud platforms, adaptive networking, data analytics and the Internet of Things.

Technology and infrastructure availability is enabling fundamental research in biomedicine, applied research for new diagnostics and the development of tools to assist people with sensory processing challenges such as autism. Curtin University is participating in Western Australia's pioneering Data Linkage System and the national Population Health Research Network with the aim of bringing data-driven innovation to population health. The Curtin Health Innovation Research Institute focuses on the public health challenges of ageing populations and chronic conditions. Researchers are developing treatments for conditions including cardiovascular disease and neurological disorders, and are trialling interventions for age-associated cancers.

Looking forward

The university remains dedicated to its mantra, *make tomorrow better.* Initiatives to tackle future challenges include the Curtin Open Knowledge Initiative to explore the implications of open knowledge and access systems, and battery research to foster innovation and industry development in energy storage; the Future of Work Institute to prepare professionals for the careers of tomorrow; and the Western Australia Data Science Innovation Hub to coordinate and boost resource sector leadership in digital systems, processes and technologies. ■

Curtin University

research.curtin.edu.au

Advertiser retains sole responsibility for content

A BIODIVERSITY HOTSPOT MAKES A NATURAL LABORATORY

THE UNIVERSITY OF WESTERN AUSTRALIA taps into one of the world's richest reservoirs of resources to develop exceptional research programmes

Access to a wealth of natural and historical resources, combined with world-class research facilities and inter-disciplinary teamwork, make The University of Western Australia (UWA) an exciting place for researchers.

Studying the seas

The 21,000 kilometre Western Australian coastline stretches from the tropical Kimberley region to Australia's temperate south coast. UWA is involved in a number of research projects studying and exploring these Indian Ocean waters.

UWA recently opened its Wave Energy Research Centre in Albany. Research underway at the centre will increase our knowledge and understanding of wave, tidal and offshore wind energy, and put the state at the forefront of marine renewable energy technology.

The Indian Ocean Marine Research Centre (IOMRC) is a collaboration between UWA's Oceans Institute, CSIRO (Commonwealth Scientific and Industrial Research Organisation), the Australian Institute of Marine Science and the Department of Primary Industries and Regional Development's fisheries division. The IOMRC has a purpose-built six-storey facility on UWA's Perth campus and a refurbished marine station on the shore of Waterman's Bay. These facilities represent an AUD$73 million investment in biochemistry, hydrodynamics, oceanography and computer modelling. They allow more than 300 researchers in multidisciplinary teams to focus on pressing issues facing the world today: climate change, sustainable use of marine resources, conservation of marine biodiversity, coastal zone management, and marine security and safety.

A treasure trove of plants

It's not only the marine environment attracting researchers. Western Australia's southwest corner, an area roughly the size of England, is one of only 36 named biologically rich zones in the world. "No comparable area on Earth can match the age of discovery witnessed here recently," says Stephen Hopper, professor of biodiversity at UWA's Centre of Excellence in Natural Resource Management. The area's unique flora and fauna provide unmatched opportunities for research. It hosts the world's highest amount of plants pollinated by birds and mammals (15%), and more than 4,000 species of plants are endemic to the region. Hopper has named 300 plants in the area, new to science, including eucalypts, orchids and kangaroo paws.

FEW LANDSCAPES OFFER AS MUCH TANGIBLE EVIDENCE OF HUMAN HISTORY

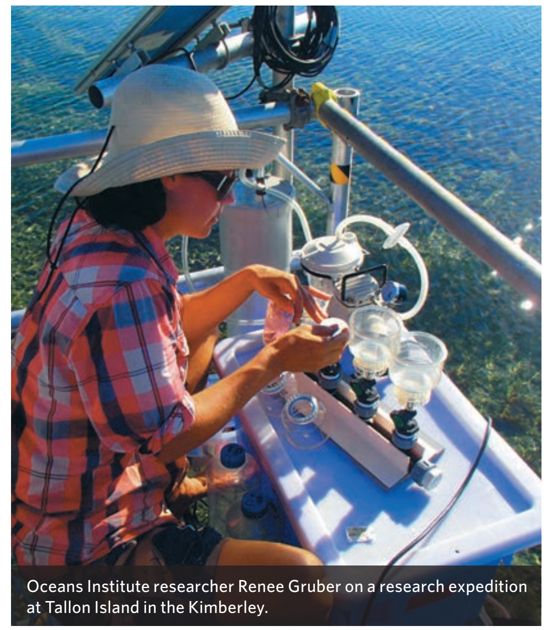
Multidisciplinary approach to evolutionary biology

Researchers at the Centre for Evolutionary Biology are tackling some of evolution's deepest mysteries. "Our research addresses evolutionary questions in diverse organisms, from insects to humans," associate professor Amanda Ridley explains. The centre adopts a multidisciplinary approach to explore selective processes acting on the morphological and life-history traits of whole organisms and their gametes. It includes experts in acoustic signalling, predator–prey interactions, visual ecology, sperm competition, chemical ecology and genetic mapping of complex traits.

"Our work is carried out both in the field and the lab, in natural and artificial settings," says Ridley. "We're highly collaborative at both national and international levels and combine skills across the board from a whole-organism approach, to intricate detail at the cellular level. In a nutshell, we robustly test scientific principles and challenge the traditional ethos of what science predicts we should see."

UWA researcher Ben Ashton testing the cognitive ability (and weight) of magpies.

Professor Stephen Hopper in the southwest region he has so much passion for.



Oceans Institute researcher Renee Gruber on a research expedition at Tallon Island in the Kimberley.

Bird track engravings looking over Deep Gorge in Western Australia's north west.

© Matthew Galligan

The team's work is frequently reported in the media, which includes a spotlight on their recent research into the cognitive ability of magpies. The range of research underway and the chance to work in the wild make the centre a popular option for PhD students.

Rock canvases

UWA's Centre for Rock Art Research and Management digs back through time, tapping into some of Australia's most spectacular rock art galleries. "Few landscapes offer as much tangible evidence of human history as the Pilbara, Kimberley and Western Desert regions in Western Australia," says the centre's director, Jo McDonald. "They present archaeologists and rock art researchers with an extraordinary opportunity to learn more about the rich visual histories associated with them."

The Burrup Peninsula, or Murujuga, on the mid-west coast of Western Australia is one such place, home to over one million indigenous engravings on piles of ancient blocks. "Some of this art demonstrates the first use of this land by people arriving over 45,000 years ago, when the hills were more than 100 kilometres from the coast," McDonald says.

A leader in the region

With such a broad range of unique natural resources in Western Australia, UWA is perfectly placed to lead research in a number of key areas. "We're regarded as one of Australia's top institutions, attracting people of world standing across a number of fields, including astronomy, marine science, the geosciences, agriculture, biodiversity, climate change and health," says Robyn Owens, deputy vice-chancellor for research. ■

THE UNIVERSITY OF WESTERN AUSTRALIA

UWA contact
Deputy Vice-Chancellor (Research)
Professor Robyn Owens
Tel: +61 8 6488 2460
E-mail: dvcr@uwa.edu.au.

Advertiser retains sole responsibility for content

The young gun of Western Australia

Though a young institution, EDITH COWAN UNIVERSITY HAS RAPIDLY TRANSFORMED from purely teaching beginnings to applied research that is making a difference.

Enabling early detection of melanomas; enhancing cyber security; and exploring the role seagrass plays in climate change are some of the vital areas being investigated at Western Australia's Edith Cowan University (ECU), whose growing reputation for specialized research and infrastructure belies its status as the state's youngest higher education institution. Such high-impact research has led to ECU being rated in the top 150 universities less than 50 years old globally by the Times Higher Education Young Universities Rankings.

In partnership with Perth hospitals, clinics and health institutes, ECU is developing a blood test for detecting early stage melanoma. The Melanoma Research Group discovered it was possible to detect autoantibodies produced when a melanoma first develops. It is crucial to detect melanomas as early as possible in order to provide the best patient outcomes. Currently 75% of biopsies are negative, whereas the blood test offers 80% accuracy. The blood test will provide clinicians with an additional tool for improving diagnostic reliability in cases of amelanotic melanomas and very early thin melanomas for patients with a family history of melanoma. ECU's deputy vice-chancellor of research, Caroline Finch, says the melanoma study is fundamental work with an applied focus that relies on collaboration with medical practitioners.

In the rapidly evolving arena of cyber security, ECU is working with Interpol, as well as state and federal police and legal authorities as part of the cutting-edge ECU Security Research Institute. Estimates suggest the rise of cyber crime will result in a global shortage of 1.5 million security professionals by 2020. ECU is the base of the Cyber Security Cooperative Research Centre, and one of just two Australian institutions considered an Academic Centre of Cyber Security Excellence.

Again, collaboration is the key — ECU is known as a leader in cyber security through interacting closely with industry partners. "We work with a lot of the key companies in the area, with the Australian Department of Defence and governments — places where cyber security is critical and that are at the forefront of being able to identify threats," Finch says. "We have the highest level of security clearance available to an Australian university, which enables us to help with Department of Defence work."

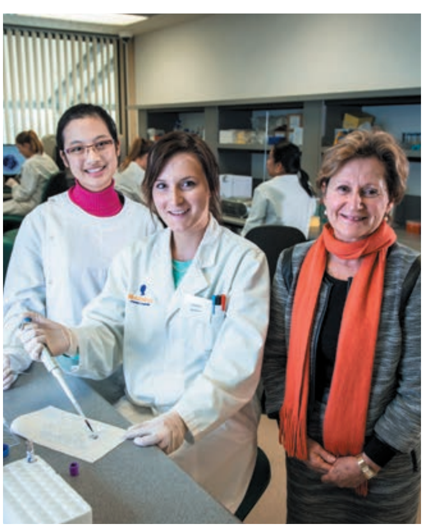
WE HAVE THE HIGHEST LEVEL OF SECURITY CLEARANCE AVAILABLE TO AN AUSTRALIAN UNIVERSITY

ECU researchers have also worked with scientists from other institutions to establish that seagrass die-off due to a summer heatwave could have increased Australia's annual carbon dioxide emissions from land use by 21%, indicating an area that needs to be a focus of mitigation efforts.

"Our goal is to capitalize on our research infrastructure and resources and then to both drive and support government and industry action in response to our findings," Finch explains. ■

AUSTRALIA ECU EDITH COWAN UNIVERSITY

www.ecu.edu.au/research



Making a difference through personalized medicine

Researchers from WESTERN AUSTRALIA'S MURDOCH UNIVERSITY are transforming people's lives and having global impact through precision medicine programmes.

Murdoch University researchers' rise as world leaders in personalized medicine is dramatically illustrated by the story of Billy Ellsworth. Billy suffers from Duchenne muscular dystrophy (DMD), a fatal childhood condition afflicting boys, requiring them to use a wheelchair by age 12.

When he was 10, Billy was on that track — his breathing was erratic and he was unable to walk unaided on small inclines. But then Steve Wilton and Sue Fletcher, researchers who are now both at Murdoch University, designed a new drug aimed at helping people with DMD, called Exondys 51.

After taking Exondys 51 for two years, Billy's breathing had stabilized and he was able to walk up inclines independently, while whistling! Now 17 and still on the drug, he remains on his feet.

Exondys 51 delayed the loss of muscle function and reduced disease severity by restoring dystrophin, a protein that is missing in DMD sufferers. In late 2016, it was accelerated for approval by the US Food and Drug Administration.

"Despite being a first-generation drug, it has altered disease progression," says Wilton. "We're hearing some really positive stories of people responding to continued treatment with the drug."

Exondys 51 is just one example of the ground-breaking work being done by Murdoch University researchers on personalized medicine for rare and infectious diseases.

WE'RE HEARING POSITIVE STORIES OF PEOPLE RESPONDING TO THE DRUG

Elizabeth Phillips is a professor at both Murdoch University and Vanderbilt University Medical Center. In global collaborations, she has identified biomarkers for adverse drug reactions that might occur when patients are being treated for high burden and once deadly diseases such as HIV. The work has lead to earlier diagnosis, treatment and prevention.

Between 2002 and 2008, Phillips, Simon Mallal and their team at Murdoch discovered a strong association between an adverse drug reaction to the HIV medication abacavir and the genetic marker HLA-B*57:01. Their championing of genetic screening in routine HIV clinical practice led to the first multicentre randomized clinical trial to show a specific genetic marker can be used to prevent abacavir hypersensitivity. This triumph in personalized medicine has now led to the prevention of drug hypersensitivity in thousands of patients, creating a roadmap from discovery to translation which Phillips and others are continuing to apply to make drugs safer globally.

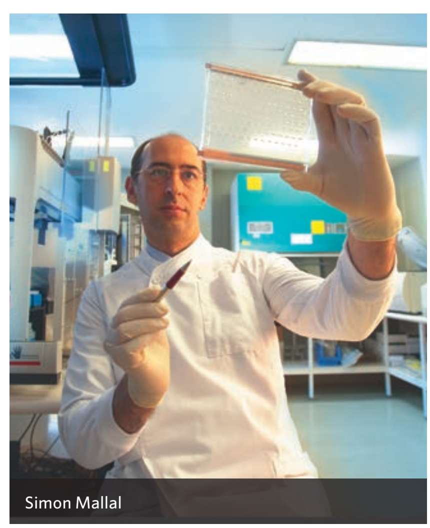
"It became very much a story of an enormous international collaborative effort," she says. "Not just between scientists and clinicians, but also with industry, to turn discoveries into tests used in clinical settings." ■

Murdoch University

www.murdoch.edu.au


Stephen Wilton and Sue Fletcher

Professor Elizabeth Phillips


Simon Mallal

Collaborative research for community benefit

THE UNIVERSITY OF NOTRE DAME AUSTRALIA has gained a reputation for ground-breaking collaborative research designed to help those facing challenging circumstances.


Researchers Lynley Wallis (L) and Anna Dwyer (R) from Notre Dame University's Nulungu Research Institute in Western Australia.

The University of Notre Dame Australia recognises the importance of collaborative research and with the foundations of Catholic intellectual tradition, Notre Dame researchers across the campuses in Fremantle, Broome and Sydney work to deliver practical benefit to communities and lives.

Few projects better illustrate the positive impact of Notre Dame's research on communities in need than the seminal investigation into alcohol restrictions in northern Western Australia. Following 2007 evidence that alcohol misuse was linked to violence and poor health and education outcomes in remote communities, the local government restricted sale of full-strength alcoholic drinks. Notre Dame researchers and Nulungu Research Institute staff working with public agencies, other universities and Aboriginal cultural and community leaders sought to gain accurate information on the effects of the restrictions. Researchers found there was more than 30% fewer alcohol-related illnesses and crimes in the first year of restrictions. Continued work demonstrates continuing declines in domestic violence and street drinking as well as improvements in family health awareness.

Notre Dame is particularly renowned for its health and medical research underpinned by strong partnerships. Research has improved survival outcomes for people with prostate, colorectal, lymphoma and sarcoma cancers through the development of a programme focused on patient needs from the time of diagnosis to beyond the completion of treatment. Other peer-led programmes significantly reduced falls among older men and women in community and residential care facilities, and new education programmes reduced the number of clinical interventions in childbirth.

RESEARCH HAS IMPROVED SURVIVAL OUTCOMES FOR PEOPLE WITH PROSTATE, COLORECTAL, LYMPHOMA AND SARCOMA CANCERS

The university is also participating in an Australia-wide industry-supported project aimed at improving detection and management of the genetic condition, familial hypercholesterolaemia, by transferring responsibility from the tertiary hospital sector to the less expensive and more easily accessible primary health care sector.

Working with investigators from the University of Southern Queensland, James Cook University and Northern Archaeology Consultancies, Notre Dame researchers are studying historical records to reconstruct a picture of 200 campsites used by the Queensland Native Mounted Police (Aboriginal troopers enlisted in the early days of European settlement). The work, funded by the Australian Research Council, is enabling a better understanding of early Australian frontier conflict and contributing to global studies of Indigenous responses to colonialism.

Notre Dame is committed to making a real difference in the communities it serves. Along with health and Indigenous studies, its primary areas of focus are ethics, philosophy, theology and education. ■

HUMAN OCCUPATION OF NORTHERN AUSTRALIA BY 65,000 YEARS AGO

Chris Clarkson, Zenobia Jacobs, Ben Marwick, Richard Fullagar, Lynley Wallis et al.

Notre Dame's focus on Indigenous studies was reflected in an article published in Nature in 2017. The article achieved a social impact score of 1425.

Published: Nature 2017: Volume 547(7663):303–310.

Clinical Sciences
'Well above world standard'

State of Australian University Research 2015-16: Volume 1 Excellence in Research for Australia (ERA) National Report

THE UNIVERSITY OF NOTRE DAME AUSTRALIA

notredame.edu.au

Published in print as a Sponsor Feature in *Nature* 562, 7727 (18 October 2018)

How 'the crowd' is tackling a silent killer

Citizen scientists around the globe are uniting to make **AFLATOXIN** a thing of the past.

AUTHORS
Howard-Yana Shapiro,
Chief Agricultural Officer,
Mars, Incorporated

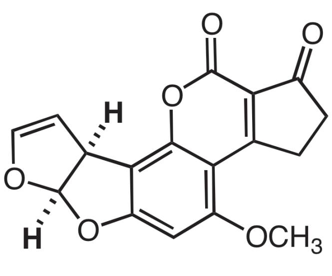
Justin B. Siegel,
Associate Professor, Department of Biochemistry, Chemistry, and the Genome Centre, UC Davis

Aspergillus flavus **– the fungus that produces aflatoxin – growing on corn.**

AFLATOXIN, A SILENT KILLER IMPACTING PEOPLE AND CROPS AROUND THE WORLD

It is difficult to believe, but 4.5 billion people globally are chronically exposed to a harmful group 1 carcinogen through their food[1], yet they're either unaware or without alternative options. This little known, wide reaching poison is called aflatoxin. The health impacts of aflatoxin are staggering: it is the most potent naturally occurring liver carcinogen we know, and is estimated to play a part in up to 28% of liver cancer cases globally[2]. Furthermore, consuming aflatoxin-contaminated food is associated with stunting in children, damage to the immune system, maternal anaemia and mortality.

Several approaches for managing and degrading aflatoxins are currently in practice, but none are widely considered to be effective. Scientists think an enzyme can be created to attack and degrade aflatoxin, which would decrease toxicity by several orders of magnitude. As a result, a group of uncommon collaborators including Mars, Incorporated, the University of California, Davis, Thermo Fisher Scientific, the University of Washington, Northeastern University, the Partnership for Aflatoxin Control in Africa (PACA) and the United Nations Food and Agriculture Organization (UN FAO) have come together to tap into the power of online, gamified protein folding with hopes that it can expedite progress towards a solution.



Chemical structure of aflatoxin B_1.

Aflatoxins are a type of mycotoxin – poisonous natural products made by certain fungi that can evoke a toxic response when consumed, even in low concentrations. These mycotoxins are largely invisible compounds that can grow in or on almost all grains and groundnuts. It's estimated that mycotoxins contaminate at least one quarter of food crops around the world, such as maize, tree nuts, cassava, millet, peanuts, several spices and even animal feeds[3]. Under certain circumstances, the toxin can contaminate crops in the field and in storage, with fungi growth most pronounced post-harvest, especially in moist, warm conditions.

In mature economies, expensive monitoring and advanced food safety technologies track aflatoxin levels to keep it out of the food chain. Where the toxin level is above legal limits, the food goods are disposed of. Although total food loss and waste figures are not known, it is widely understood that aflatoxin infestation causes a significant volume of food waste globally.

Aflatoxin disproportionately affects people in poorer countries. Even though food aflatoxin limits are set, they often go unenforced partly due to under-regulated food manufacturing and reduced access to monitoring technology. As a result, the impacts of aflatoxin are felt strongly in poorer countries.

THE POWER OF PROTEIN FOLDING

Proteins are the molecular building blocks for almost every process in living things and are responsible for structure, function and regulation in our bodies. Whilst deoxyribonucleic acid (DNA) is the posterchild of biology and provides the instructions for life in the form of our genetic code, proteins are essential in mobilizing this genetic information so that organisms can develop, reproduce and live. In the past few decades, biotechnology has empowered scientists to remodel fundamental molecules through a process called 'protein folding' in an effort to solve a range of medical challenges, such as diabetes, arthritis and cancer.

Proteins are made up of strings of amino acids that determine a protein's unique three-dimensional (3D) structure, and can be rearranged into different shapes. In fact, there are so many possibilities for potential proteins that it makes the quantity of stars in the universe look tiny. The unique structure of each protein dictates its function, and so scientists can aim to change a protein's behaviour by rearranging its structure.

Scientists and medical doctors recognize the tremendous opportunity that protein folding could offer to treat many medical conditions. Whilst most scientific discoveries in this realm typically take place in private labs, some scientists have identified the

Published in print as a Sponsor Feature in *Nature* 562, 7727 (18 October 2018)

An example of the image that a gamer will see when folding proteins on the Foldit platform.

opportunity to leverage the power of 'the crowd' to explore new protein structures at an expedited rate. One such example is Foldit, an online science crowdsourcing platform that can be played by anyone with a computer and an imagination. It challenges participants to play with 3D protein puzzles that scientists can use to tackle real world medical challenges. Launched in 2008 by David Baker and scientists from the University of Washington, Foldit has already resulted in multiple crowdsourced breakthroughs. For example, in only three weeks gamers solved the structure of an enzyme involved in the reproduction of human immunodeficiency virus (HIV) – a mystery that had stumped scientists for more than a decade[4]. In the past five years, scientists at the University of Washington's Baker Lab have used protein structures fashioned by Foldit participants around the world that act as biosynthetic catalysts[5], fight coeliac disease[6] and treat anthrax infections[7].

The process of digital protein folding can be modelled using computer algorithms and software, thanks to advancements in the field of computational biology pioneered by Michael Levitt, Martin Karplus and Arieh Warshel, winning them the Nobel Prize in Chemistry in 2013.

But why can't computers solve the protein folding challenges on their own? Classical computers cannot sample all 3D orientations at once, because this requires enormous amounts of processing power. Classic protein modelling software is programmed to try combinations essentially at random – named the 'random-walk' approach. There are of course supercomputers, some of which are specifically built for protein modelling – take IBM's Blue Gene or D.E. Shaw's Anton as examples – yet these are still unable to sample sufficient structural space to be effective in searching the designable protein landscape. With essentially infinite possible protein orientations and thousands of medical challenges, using random-walk algorithms on supercomputers doesn't quite cut it. Crowdsourcing offers a pathway to investigate new structural possibilities beyond what is traditionally explored.

HUMAN INTUITION AT AN UNPRECEDENTED SCALE

Since the Foldit Aflatoxin Challenge launched on United Nations' World Food Day 2017, gamers from across the globe have been competing to design enzymes that can tackle aflatoxin. In less than one year, players have designed more than 1.6 million models to potentially degrade aflatoxin. The time these players have taken to morph the 3D molecules equates to approximately 80,000 player hours – that's equivalent to the labour force of over 100 full-time employees working on the problem for one year.

The gamers are diverse – ranging from 12 to 80 years old. What connects them is the innate spatial reasoning abilities of the human brain, enabling gamers to tackle some of the hardest problems in biology today. Humans are geared to detect patterns in everything they experience, and Foldit takes advantage of this puzzle-solving intuition. In fact, within the community of aflatoxin gamers, some of the highest scoring Foldit accounts are owned by individuals with no scientific training. This approach provides a scale and diversity of protein-folding capabilities that cannot be achieved by an individual lab.

HOW THE FOLDIT AFLATOXIN COLLABORATION WAS ESTABLISHED

Mars believes that food safety is fundamental to food security. Food security is defined by the UN FAO as all people at all times having access to sufficient, safe and nutritious food. Because Mars is a global business, it recognized the problem of aflatoxin and its impact on food supply around the world. In India, the company was rejecting up to 70% of the peanuts at the factory gate because of elevated levels of aflatoxin. There was no guarantee, however, that the rejected crops were not entering the food supply chain elsewhere. To address this little known but devastating food safety issue, a group of uncommon collaborators came together to tackle this challenge.

On UN World Food Day 2017, a unique partnership was born between University of California, Davis, Thermo Fisher Scientific, the University of Washington, Northeastern University, the Partnership for Aflatoxin Control in Africa (PACA), the United Nations Food and Agriculture Organization (UN FAO) and Mars, Incorporated. The group recognized the opportunity to use the Foldit gaming platform to leverage the scientific capabilities of the partnering scientists and the skills and imaginations of gamers globally.

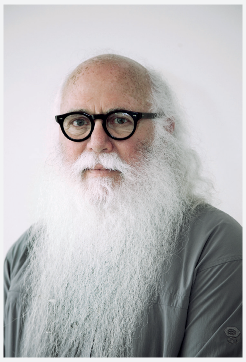
If a solution to taking down aflatoxin is produced from this collaboration, it will be freely available to anyone in the world. Mars and partners have committed that all player designs will be available in the public domain, free from patents, to maximize the positive impact that this project could have on global food safety.

SYNTHETIC BIOLOGY AS A FORCE FOR FOOD SAFETY

Over the past year there have been 12 gaming rounds of the aflatoxin puzzle released on the Foldit platform. After each game round, the best scoring models are picked for the next round of the process – analysis. Scientists at the Siegel Lab at the University of California, Davis analyze the protein structures for their amino acid sequence and send the information to Thermo Fisher Scientific. Thanks to advances in synthetic biology technologies, the amino acid information is translated and optimized into biology's digital code – DNA, the software of life. DNA is physically produced by Thermo Fisher Scientific's synthetic

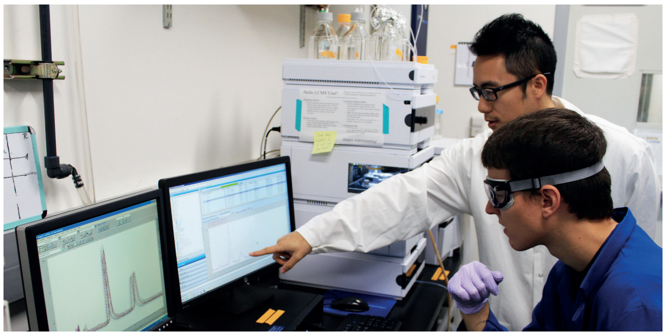
A young scientist playing the Foldit Aflatoxin Challenge at the 68th Lindau Nobel Laureate Meeting.

Published in print as a Sponsor Feature in *Nature* 562, 7727 (18 October 2018)



DR. HOWARD-YANA SHAPIRO, CHIEF AGRICULTURAL OFFICER, MARS, INCORPORATED

There isn't time to wait to solve the problems that aflatoxin creates, because it's already impacting the health and well-being of 4.5 billion people. At Mars we want to improve food safety and security for people around the world, including the most in need and low-income populations.



Researchers in the Siegel Lab at UC Davis, Wai Shun Mak (pictured above in background) and Ryan Caster (project lead, pictured above in foreground), analyzing proteins for aflatoxin activity.

biology team, leveraging their proprietary oligo and gene synthesis capabilities that encode for the newly designed proteins. Future DNA synthesis runs will use Thermo Fisher's miniaturized semiconductor-based nucleic acid synthesis platform. This technology can generate 35,000 individually selectable oligos manufactured at the same time, which are then stitched together to make up the code for the enzyme.

Once created, the synthesized DNA molecules are sent back to the Siegel Lab to see if – when expressed and folded into real proteins – they have the ability to detoxify aflatoxin. The scientists are particularly interested in targeting aflatoxin's susceptible lactone ring, because this allows the synthesized DNA to form an enzyme capable of performing lactone hydrolysis on aflatoxin B_1 under industrial conditions where only water needs to be present to perform the desired chemistry[8]. Chemical degradation of this lactone ring through enzymatic hydrolysis has the potential to decrease aflatoxin mutagenicity by more than 400-fold[9].

THE ROAD AHEAD

Although there is still much work to be done, scientists are already seeing forward momentum. With more promising amino acid strings produced with every iteration, it is only a matter of time before synthetic biology produces an enzyme that can do the job.

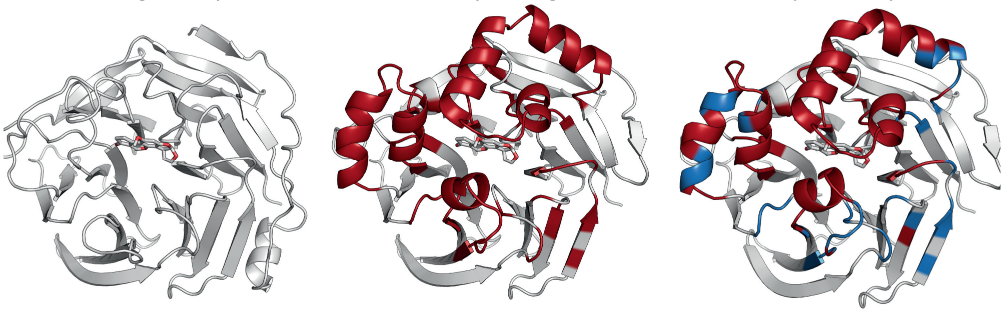
The Foldit platform not only provides the protein designs, but facilitates knowledge exchange between citizen scientists and the researchers behind the scenes. What is promising is the way in which gaming rounds learn; they fine tune the digital protein to enable gamers to reach more meaningful 3D orientations. There exists a symbiosis between the Foldit gamers and the Siegel Lab, because each new round builds off feedback from the last to optimize the puzzle and the solution selection parameters. Recent rounds have allowed gamers greater movement of the aflatoxin molecule and trimmed the scaffold protein area, enabling Foldit players to design an active site that creates a more tightly bound molecule.

As the project continues to evolve, more and more game players are taking part in the movement. This is not only an effort to neutralize aflatoxin, it is also a step towards inclusive citizen science and open dialogue between a diverse set of collaborators.

What we are seeing is a new approach to one of society's grand challenges. The partners are tackling one of the most persistent carcinogens in a way that has not been attempted since aflatoxin was discovered in the 1960s. If efforts are successful, the positive impacts will largely be felt in young economies, where aflatoxin unfairly burdens society. Promisingly, the Foldit Aflatoxin Challenge has brought together a unique set of collaborators – industry, academia and the general public – to advance scientific discovery and, ultimately, tackle the persistent threat of aflatoxin around the world.

REFERENCES

1 Williams, J. H. *et al. Am. J. Clin. Nutr.* **80**, 1106–1122 (2004).
2 Liu, Y. & Wu, F. *Environ. Health Perspect.* **118**, 818–824 (2010).
3 Marin, S., Ramos, A. J., Cano-Sancho, G. & Sanchis, V. Food Chem. Toxicol. **60**, 218–237 (2013).
4 Khatib, F. *et al. Nature Structural & Molecular Biology* **18**, 1175–1177 (2011).
5 Eiben, C. B. *et al.*.*Nature Biotechnology* **30**, 190–192 (2012).
6 Gordon, S. R. *et al. J. Am. Chem. Soc.* **134**, 20513–20520 (2012).
7 Wu, S. J. *et al. J. Biol. Chem.* **286**, 32586–32592 (2011).
8 Nicolás-Vázquez, I., Méndez-Albores, A., Moreno-Martínez, E., Miranda, R. & Castro, M..*Arch. Environ. Contam. Toxicol.* **59**, 393–406 (2010).
9 Ehrlich, K.C., Moore, G. G., Mellon, J. E. & Bhatnagar, D. *World Mycotoxin J.* **8**, 225–233 (2015).



The stages of player protein designs on the Foldit platform. (Left: original enzyme. Centre: enzyme with gamer design. Right: enzyme with player design and computational refinements).

African students at the Lindau Nobel Laureate Meetings: At the Africa Outreach breakfast with Nobel Laureate Peter Agre (left) and during discussions

Lindau Nobel Laureate Meetings

Africa's Next Generation

How to Support Africa's Science Structures for Tomorrow's Best Young Scientists / by Wolfgang Huang and Stefan Kaufmann

Every year, the best young talents in sciences gather in Lindau, Germany, to meet with Nobel Laureates as well as their peers. 600 students and post-docs are chosen in a multi-stage selection process from all around the world for this once-in-a-lifetime experience. As participation from Africa was lagging behind for many years, the Africa Outreach Initiative under the patronage of former German Federal President Horst Köhler was started in 2015. Supported by the Robert Bosch Stiftung, the initiative helped to bring more than 150 excellent young African scientists to Lindau. In 2019, South Africa will be hosting the meeting's "International Day" – a highlight of Africa's presence at the Lindau Nobel Laureate Meetings.

With this in mind, we asked several Lindau Alumni as well as partners of this initiative about their thoughts on the current status of scientific excellence in Africa, what progress has been made, and what still needs to change.

A Decade of Progress

While the situation greatly varies among African countries, the last decade has seen a considerable growth of scientific agencies, programmes, networks and conferences, and certainly an improvement of the situation.

To no one's surprise, South Africa is spearheading this development with it's National Research Foundation, established almost 20 years ago. Current programmes such as the South African Research Chairs Initiative and the Centres of Excellence funding scheme contribute to keeping excellent scientists in Africa, says Roseanne Diab, Executive Officer of the Academy of Science of South Africa (ASSAf). But she also highlights various cross country-intiatives: "The African Institute for Mathematical Sciences (AIMS) is a pan-African network of centres of excellence for postgraduate education, research, and outreach in mathematical sciences established in 2003. This was followed more recently by the AIMS Next Einstein Initiative, the goal of which is to build fifteen centres of excellence across Africa by 2023."

Most progress has been made in the area of health; all the more important as a bad public health situation has countless negative efffects on people, economies and countries – and on science.

There is a heroic effort to meet the Sustainbable Development Goal #3 by 2030.

Berhanu Abegaz

For example, between 2000 and 2015, the number of malaria deaths has decreased from an approximated 839,000 to 438,000: a decline by 48%. 90% of malaria infections occur in African countries.

SDG 3: "Ensure healthy lives and promote wellbeing for all at all ages"

Challenging Structures and Attitudes

Yet, only 1% of global investment in R&D is spent in Africa, and the continent holds a tiny 0.1% share of the world's patents, as ASSAf's liaison officer Edith Shikumo points out. But money doesn't seem to make up the top priority on younger scientists' list of concerns. "I don't want to mention the usual obstacles like lack of proper infrastructure and expensive equipment; I would rather focus on the lack of tolerance for new and innovative ideas, the fear associated to out-of-the-box thinking and the tendency to avoid risk accompanying entrepreneurship are the main obstacles for a thriving science and research culture", says Ghada Bassioni, guest professor at the Technische Universität München and coordinator of Egypt-Germany collaboration, with the Science and Technology Development Fund of the Ministry of Higher Education and Scientific Research, Egypt – and Lindau Alumna.

Mark Williams-Wynn, who attended the Lindau Meetings in 2016 and is currently a post-doctoral fellow at the University of KwaZulu-Natal, adds: "One of the biggest obstacles that I have noticed is the tendency to avoid questioning the status quo. There is an attitude which exists not only among academic institutions, but throughout society: 'if it isn't broken, why fix it'. This applies not only to innovation, but also to how people go about their work."

There is a lack of a desire to excel, and the aim of many people is to fulfil only the minimum requirements.

Mark Williams-Wynn

"This lack to excel creates an obstacle to a thriving research culture, as not only is there a lack of innovation to be found, but there is little to no support for the commercialisation or further development of innovation."

But as Lydia Rhyman, affiliated at the University of Mauritius and the University of Johannesburg and a 2017 Lindau Alumna, points out, there is still more needed to create a thriving science environment, such as: (1) a mature, supportive leadership culture, (2) increased collaborative efforts, particulary south-south cooperations, (3) better access to infrastructures and equipment, (4) less PhD brain drain, (5) political and economic stability.

Career Paths for Young Scientists

When asked about the availabilty of suitable, reliable career paths, everyone agrees that this is one of the major problems that limits progress and success of African research: Excellent young African PhD students are lured away by better conditions, more exciting science and more money, and due to the research structure situation, they may not be coming back. Of those who return, a large number chooses South Africa, where conditions are still the best. Plus, many are absorbed by industry, where they are lost from research fields, as Williams-Wynn observes.

Pursuing a career in research can be daunting in Africa. Investment in young scientists is required so that Africa can come up with innovative solutions for its development.

Lydia Rhyman

But there is hope, says Berhanu Abegaz, former executive director of the African Academy of Sciences: "There are many networks and organisations that are now available to help young people and to get them focused on African issues and also get them to stay in Africa." These include the African Academy's Affiliates program, the Next Einstein Forum, the AIMS, the Africa-Oxford Initiative, the Organisation for Women in Science for the Developing World, the African Women in Agricultural Research and Development, leadership academies, young academies, and others. Yet only a few African countries invest enough money in such structures and programmes.

From Support to Cooperations

Some of the programmes mentioned have been initiated and funded by Northern countries, mainly Europe and the United States. However, Europe's history in the colonisation of Africa and the resulting sentiment that supporting Africa means paying an outstanding debt doesn't fit to a modern approach. Abegaz clarifies that "engagements will be beneficial to all parties if the driving philosophy is equitable partnerships. This would have to begin in the agenda setting stage and defining the basis of the partnerships. Some European partners like the Wellcome Trust accept, at least in principle, this approach."

Roseanne Diab adds another example: The Developing Excellence in Leadership, Training and Science (DELTAS) Africa programme has already invested £60 million in training African researchers. Funding comes from the Wellcome Trust and the UK government, but the programme is led not from London, but from Nairobi, Kenya, by the Alliance for Accelerating Excellence in Science in Africa (AESA). By shifting the centre of gravity in this way, DELTAS is beginning to ensure that science for Africa is led by Africa's researchers, and that it remains relevant to the needs of the continent.

Downsides of foreign partners are linked to research agendas being drawn by funders that do not serve Africa's interests nor research needs. As a consequence, more South-to-South partnerships need to be encouraged for Africa's development.

We would want to see the role of other countries as equal partners and the investment in legacy projects that help to leave a tangible footprint by ensuring ownership by African countries and investments in its human capital.

Roseanne Diab

In a shift of perspective and wording, it is no longer help that African countries are requesting, but partnership, explains Rhyman as she adds: "We African scientists cannot consider ourselves to be always in a position of requesting help from others, it is high time for us to wake up and work. For how long the support will continue? Why do they need to support us? When will we stand on our feet?"

Wolfgang Huang is Director of the Executive Secretariat of the Lindau Nobel Laureate Meetings.

Stefan Kaufmann is Director at the Max Planck Institute for Infection Biology, Berlin. He was one of the scientific chairmen of the 2018 Lindau Meeting.

nature OUTLOOK

SCIENCE WITHOUT BORDERS

18 October 2018 / Vol 562 / Issue No 7727

Cover art: Taj Francis

If a researcher's goal is to improve people's lives, they should look to the world's poorer nations. Low- and middle-income countries — customarily grouped by the broad and flawed umbrella term 'the developing world' — are where the greatest gains can be found.

These nations commonly lack the wealth and resources needed to access and act on the world's scientific knowledge. What they do not want for is talent. And with more than half of the world's population growth between now and 2050 expected to come from low- and middle-income countries, it is more pressing than ever that their young people realize their talent.

As a country's population grows, the problems that it faces change. Non-communicable diseases such as diabetes are placing an increasing burden on health systems. In countries as far removed as Peru and India, the capacity to address these new priorities is being built (see page S65).

One threat to capacity-building in emerging economies is brain drain. Many budding researchers leave Africa to study in Europe and North America, and not all come back. And those who do return often find themselves unprepared for the challenges that await them (S58).

Not all the hurdles young researchers in low- and middle-income countries must overcome are unique to those locales. From the 68th Lindau Nobel Laureate Meeting in Germany this June came a call for a global commitment to support science (S64). Those assembled also discussed the central importance of researchers maintaining the trust of the communities in which they work (S62). Few Nobel prizes for science have gone to researchers from low-income nations. It falls on these countries' young researchers, and anyone with the power to assist them, to redress the balance.

We are pleased to acknowledge the financial support of Mars, Incorporated in producing this Outlook. As always, *Nature* has sole responsibility for all editorial content.

Richard Hodson
Supplements editor

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at go.nature.com/e4dwzw

CITING THE OUTLOOK
Cite as a supplement to *Nature*, for example, *Nature* **Vol. XXX**, No. XXXX Suppl., Sxx–Sxx (2018).

VISIT THE OUTLOOK ONLINE
The *Nature Outlook Science without borders* supplement can be found at www.nature.com/collections/science-without-borders-outlook. It features all newly commissioned content as well as a selection of relevant previously published material that is made freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES
Site licences (www.nature.com/libraries/site_licences): Americas, institutions@natureny.com; Asia-Pacific, http://nature.asia/jp-contact; Australia/New Zealand, nature@macmillan.com.au; Europe/ROW, institutions@nature.com; India, npgindia@nature.com. Personal subscriptions: UK/Europe/ROW, subscriptions@nature.com; USA/Canada/Latin America, subscriptions@us.nature.com; Japan, http://nature.asia/jp-contact; China, http://nature.asia/china-subscribe; Korea, www.natureasia.com/ko-kr/subscribe.

CUSTOMER SERVICES
Feedback@nature.com

CONTENTS

RELATED ARTICLES

COMING HOME

Budding African scientists often travel to other countries for study. But how well does foreign training prepare them for the realities of science on their return?

BY LINDA NORDLING

TAJ FRANCIS

When Yaw Bediako returned to Africa after receiving high-school, undergraduate and PhD education in the United States, the Ghanaian was struck by how different life as a scientist was on the continent of his birth. As a postdoctoral researcher in Kenya working on malaria, Bediako was asked to weigh in on decisions from whom to hire to what equipment to buy. “Postdocs in the United States don’t have to think beyond the lab. It’s your principal investigator’s job to stress about funding and facilities,” he says.

The immunologist also found that African researchers have to think about their next step earlier than their counterparts in Europe or the United States do. In Africa, he says, postdoc posts tend to be short — around two years. “If you don’t get a grant during that time, there is no safety net.”

Every year, thousands of African graduates leave to pursue higher degrees abroad. Their reasons vary. For some, it is the limited training opportunities at home and the chance to learn from the best in their fields, while others see better career prospects abroad. Many return, hoping to help address the needs on the continent that they are so keenly aware of. But the journey home can be surprisingly difficult. Returnees have to contend with masses of red tape and teaching demands — more than if they had stayed away — and some struggle to find a scholarly space that can accommodate their expertise. Many give up — a loss to the continent, which needs one million new PhDs just to match the world’s researcher-to-population average. But in recent years, dedicated support programmes have sprung up to help African researchers to make the transition and return home.

“I HAD NEARLY ZERO MENTORSHIP. I HAD TO FIGURE OUT EVERYTHING ALMOST ON MY OWN.”

THE RIGHT CHOICES

For many who struggle on their return, the crux of their problems lies in not having enough help in their early careers to make informed choices. Connie Nshemereirwe is one such researcher. Growing up in Uganda, she always knew she wanted to make a difference to the people in her country. After earning an undergraduate engineering degree at home, she worked in the construction industry before discovering a passion for teaching at one of the country’s small private universities. Looking for opportunities to study further, she found a fully funded master’s degree programme in educational and training systems design in the Netherlands. It was here that she came across the field of educational measurement, which uses statistical modelling to evaluate the performance of education systems.

After a time back in Uganda, Nshemereirwe returned to the Netherlands for a PhD. For her doctorate, she wanted to address a problem back home. From experience, she suspected that the Ugandan high-school system, in which students gain access to university through national examinations, discriminates against students from poor backgrounds. Her PhD thesis, published in 2014, used mathematical models to show that this was true. Returning to Uganda, she hoped that her research could help to reform university admissions systems and make them fairer.

But that didn’t happen. The information she provided was not as useful to the Ugandan education ministry as she had expected. In a country like the Netherlands, she says, such data are often used to inform policy. But Uganda’s ministry of education, examinations council and universities do not have strong data-collection systems. They were ill-prepared to process and respond to her findings. Nshemereirwe’s efforts to quantify the problem were not of much use.

She also struggled to find a footing in Uganda’s research system. Nshemereirwe had tried to find a Ugandan co-supervisor for her PhD, to advise alongside her Dutch mentor. But she was able to identify only two scholars in her field in Uganda, and both had left academia. On her return, there were no researchers in her field with whom to spar and no community in which she could grow as a scientist. The specialism she had chosen just didn’t fit into any existing scholarly space, she says.

Nshemereirwe’s frustrations led her to quit university life and set out as an independent adviser, bridging the worlds of academia and education policy. Had she known earlier what problems she would run into, she would have made different choices. “But I had nearly zero mentorship,” she says. “I had to figure out everything almost on my own.”

MISSION IMPOSSIBLE

The anxiety that comes with failing to find your feet when returning from a prestigious stint abroad can cause confusion and self-doubt. Mashiko Setshedi, a gastroenterologist at Groote Schuur Hospital in Cape Town, South Africa, decided to move into research in her 30s. “I wanted to add something else to my repertoire,” she says.

After completing a master’s degree at the University of Cape Town in South Africa, she spent three years as a PhD student at Brown University in Providence, Rhode Island, studying chronic hepatitis and liver cancer. On her return to South Africa in 2012, she was keen to use her new skills to improve clinical care and was able to obtain a research fellowship to cover her salary. But establishing a research programme from scratch proved to be much harder than she expected. “It was mission impossible,” she says.

Although her PhD had given her a foundation, she lacked a local track record in research. As things began to go against her, she lost confidence in her own abilities. She was unable to attract students, and she struggled to get the clinical samples she needed for her ill-fated project. In the face of these difficulties, she retreated to the security of clinical work and teaching,

frustrated and doubting whether research had ever been for her. “You can get all the training you want overseas, but it doesn’t mean that when you come home the environment is conducive, or that you are actually prepared,” she says.

What helped Setshedi in the end was further training, specifically a postdoc in an immunology lab at the University of Oxford, UK. Armed with first-hand experience of the challenges that she would face at home, she sharpened her grant-writing skills and firmed up her research networks. She is now back in Cape Town and on track with her research ambitions. She has gained the confidence of the consultants in her unit, and has local mentors who are keen to help by offering their expertise, lab space and equipment. “I feel more competent in my skills and therefore more confident to do my job well,” she says.

A SOFTER LANDING

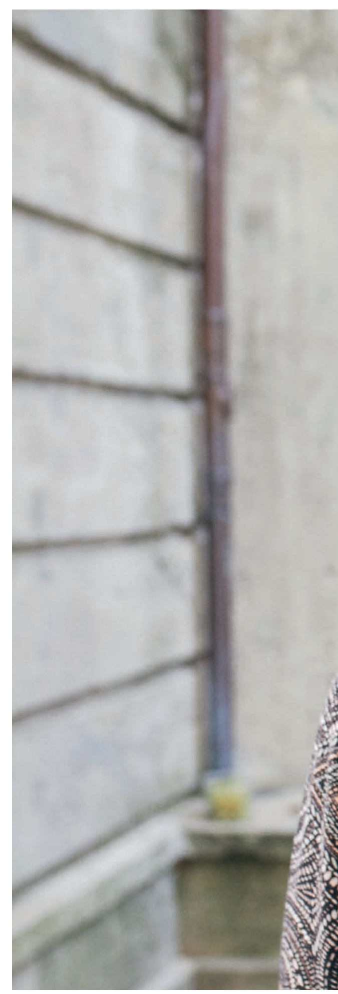
It’s not uncommon for African students to follow Setshedi’s path and do several stints abroad. Bediako followed his first postdoc in Kenya with a second at the Francis Crick Institute in London. Just as Setshedi had found necessary, he hoped to broaden his skill set and build up his scientific reputation before returning home to set up his own research group. But unlike her, Bediako found himself with plenty of academic career guidance — not least from his father, who holds doctorates in both French literature and theology. His father had also trained abroad before returning to Ghana, despite offers to continue working at foreign universities. Through his father’s circle of colleagues and friends, Bediako was exposed to a culture of proudly African academics. “Such mentorship continued throughout my academic experience abroad,” he says.

Few are so lucky. But in recent years, several support programmes have surfaced to attract back young African scientists who have moved abroad, and to help them to forge a career on the continent. For example, the Francis Crick Institute’s African Career Accelerator Awards, which were launched earlier this year, are designed to help early-career researchers to set up their own research groups. With money from the United Kingdom’s Global Challenges Research Fund, it targets African citizens who want to establish themselves as independent researchers on the continent. Fellows receive up to £130,000 (US$168,300) over two years to cover salary, research expenses and visa costs, with an additional £20,000 available for equipment purchases. It is hoped that the programme will help African scientists to make the tricky leap from postdoc to principal investigator. “It’s difficult enough to get through that period in the United States or northern Europe, but in Africa the odds are stacked against you,” says Robert Wilkinson, the scheme’s originator, who splits his time between the Crick Institute and the University of Cape Town. Bediako will be one of the first recipients.

In Kenya, the Initiative to Develop African Research Leaders (IDeAL) offers a wider gamut of training opportunities. It provides support to scientists ranging from high-school graduates to those pursuing postdocs. The initiative is run from the Kenya Medical Research Institute (KEMRI)–Wellcome Trust Research Programme in Kilifi, on the Kenyan coast. As of August this year, the initiative (which started in 2015) has supported 40 PhD students and 13 master’s students, awarded 22 postdoctoral fellowships, and produced more than 400 peer-reviewed publications.

Despite IDeAL’s largely UK-derived funding, Sam Kinyanjui, the immunologist who runs the initiative, describes it as Africa-based and Africa-driven. It is designed to train scientists to deal with the sometimes challenging research environment in Africa. Whereas the programme’s undergraduate and master’s training is done on the continent, PhD training and beyond still usually involves students spending some time

Yaw Bediako has received support from Kenya’s Initiative to Develop African Research Leaders.



BRIAN DAVID STEVENS

abroad, he says, where they hone their skills in resource-rich settings. "That sounds contrary to a lot of people," he admits. But, although IDeAL now has most of the human capacity and facilities it needs to do the training in-house, he says that there are benefits to exposing students to different environments because research is an international endeavour. The students start and finish their PhDs in Kenya and they have two supervisors throughout — one in the collaborating foreign institution, and one back home.

After submitting their PhDs, the students on the IDeAL programme can access a one-year career development grant, designed to help them find postdoc positions. It can also be used to leverage matching funding from a partner institution, creating a two-year jointly supported postdoc. "That way they retain their affiliation with us regardless of where they do their postdoc," says Kinyanjui. With the link maintained, IDeAL alumni can continue to access help, with grant applications for example, that will help them to stay in or return to Africa.

"I'D LIKE TO ESTABLISH A WORLD-CLASS IMMUNOLOGY PROGRAMME."

BEST-LAID PLANS

Bediako, who has benefited from IDeAL's postdoc support, is currently planning his return to Ghana. Early next year, he will join the West African Centre for Cell Biology of Infectious Pathogens at the University of Ghana in Accra, a centre established in 2013 by malaria researcher Gordon Awandare, another Ghanaian who returned home after many years abroad. Once there, Bediako hopes to create his own lab working on malaria immunology, then broadening to other diseases. "Ultimately, I'd like to establish a world-class immunology programme that will serve as a platform for locally driven research," he says.

Bediako thinks that all African universities should take a leaf out of IDeAL's book and do more to stay involved with their graduates who leave to train abroad. "Too often we don't have enough intellectual input into what our students are doing," he says. These universities should engage with the students before they leave, he says, such as by organizing workshops where students can talk to faculty members who have been in their shoes. And when students are abroad, their home universities should provide them with opportunities to visit and share their work and experiences. "Travel fellowships to locally organized conferences would be a good place to start," he says.

That sort of sustained contact can also help young African scientists to choose their projects and fields of study with the future in mind, so that they don't find themselves in the position of Nshemereirwe — highly skilled in a field that does not fit into the research landscape of their home country. Funding constraints in Africa are another thing to keep in mind, says Kinyanjui. Because domestic funding is scarce, scientists in Africa have to adapt their work to match the shifting research priorities of international donors — although work is under way to strengthen the African voice in setting funding priorities. To engender the kind of flexibility that may be required, his programmes expose students to different fields of study as much as possible. Every Thursday, scientists from across the KEMRI–Wellcome programme meet for a seminar on topics that range from social science to vaccines. "The whole idea is to help people see the bigger picture," says Kinyanjui.

Mankgopo Kgatle, a South African virologist who is doing a postdoc at the University of Oxford, is keen to avoid specializing in the wrong field and making life difficult for herself on her return. Her research background is in viruses that can cause cancer, such as hepatitis B, human papillomavirus and Epstein–Barr viruses. Now, she is complementing this knowledge with epigenetics — a relatively new field of research that has few experts in South Africa. Although she is aware that this could limit her ability to access funding when she returns, she hopes that by applying what she is learning at Oxford to cancer-causing viruses, which are a priority for funding in South Africa, she will be able to get the money she needs to pursue this cutting-edge field when she returns to the University of Cape Town in a few years.

Connie Nshemereirwe's training in the Netherlands was not as relevant to research in Uganda as she had hoped.

JOSUA PIORR

A SPECIAL DEDICATION

Even with all the training and assistance on offer, returning to Africa to set up a research group after years abroad takes dedication and planning. "To succeed here, you have to be the kind of person who is excited about building something new and is willing to put up with the frustration and difficulties because you want to be part of something bigger than yourself," Bediako says. This is perhaps true for scientists the world over, but Africa's particular challenges — lack of funding, old-fashioned administrative systems and poor infrastructure, to name but a few — make it many times harder. In Africa, scientists have to be entrepreneurial and self-motivated, Bediako says, "to the point of being almost crazy".

For him, the potential pay-off from coming home outweighs the difficulty he expects to face. "There are thousands like me in America or the United Kingdom," he says. "But if you come home and build something, then when your time comes to retire you can look back at having made a difference to your community, to your people. That is my motivation." ■

Linda Nordling *is a science journalist in Cape Town, South Africa.*


PHIL THUMA

Researchers at the Macha Research Trust have made strides in tackling malaria in southern Zambia.

COMMUNITIES

A matter of trust

Improving health in low- and middle-income countries requires earning the confidence of the local community, one relationship at a time.

BY LUCAS LAURSEN

Health researchers and workers use their training and the treatments available to them to prevent and treat illness. But they cannot bring their expertise to bear if they do not have the trust of the people that they are trying to treat.

This August, in the Democratic Republic of the Congo, communities in the midst of an Ebola outbreak continued traditional rural burial practices that include touching bodies, despite health workers' advice on sanitary burials. Residents in the village of Manbangu burnt down a health centre and injured an Ebola health-care worker after one resident died of Ebola. Sometimes fear and misinformation drive even more violent behaviour: in a 2014 outbreak of the disease in Guinea, residents of the village of Womey killed a group of eight visiting health workers, journalists and government officials.

A community's trust is not something that can be bought, says Peter Agre, a Nobel prize-winning chemist turned malaria researcher at the Johns Hopkins Malaria Research Institute in Baltimore, Maryland. "You have to earn trust," he said at the 68th Lindau Nobel Laureate Meeting in Germany, earlier this year.

One place that seems to have built and maintained trust over many years, Agre says, is the Macha Research Trust near Choma in southern Zambia. The centre has lowered the number of cases of malaria in the surrounding area much faster than global incidence rates have declined. It works with international partners, national and local authorities and the local community. Building up those partnerships relied on individuals who were willing to make a commitment to each other and the cause.

GLOBAL TALENT, LOCAL RELATIONSHIPS

Malaria researcher Sungano Mharakurwa had a tough decision to make in 2003 when he finished his postdoctoral fellowship at the University of Oxford, UK. He had developed new diagnostic tests for the disease, and would have been a promising candidate for a research post anywhere in the world. He wanted to go home to Zimbabwe, however, but his postdoc supervisor urged him to avoid the country, which was experiencing a general strike, political turmoil and a severe economic crash. "She was basically telling me I should not even think of going back," he recalls.

It was then that Mharakurwa saw an advertisement for a position to establish a malaria research centre at the Macha Mission Hospital in neighbouring Zambia, in collaboration with Johns Hopkins University in Baltimore, Maryland. The centre could not compete with the salary that Mharakurwa might have earned abroad. But what it could offer him was an opportunity to use his knowledge and network of international contacts on malaria's front line.

The Macha hospital was founded by Alvan and Ardys Thuma, husband-and-wife medical missionaries from the United States, in 1957. Agre is critical of certain medical missionaries and international agencies that rotate their staff too frequently to build up the long-term trust and knowledge that is needed to implement the best science-informed policies. Some

JHMRI

A. E. THUMA PHOTO COLLECTION

Peter Agre (left) says Macha Mission Hospital is an example of how trust is built over time. One of its founders (Alvan Thuma; right) worked in Africa for 20 years.

international health missions have such a high staff turnover that "human trust is stretched thin," Agre says. This was not the case for Macha. "[The Thumas] stayed there longer than many Zambians," Mharakurwa says. They established a long-standing rapport with the local community and national officials. Their son, Philip Thuma, who is also a doctor, helped to establish the research centre at Macha and is still working there. By those standards, Mharakurwa, who now has an administrative post at Africa University in Mutare, Zimbabwe, but is still doing research at Macha, is a relative newcomer, with 15 years under his belt.

The continuity of staff is just one reason why Macha's approach is respected by the local community. Several other important decisions have helped to gain and sustain trust. One is to hire locally. A local chief was resistant to the mission until his son started volunteering in the Macha labs. "He was able to bring his father to come and see the labs and the blood samples. And they loved it," Mharakurwa says. "The community, they feel that they are stakeholders. Their children are actually working at this institute; they have a sense of ownership," he adds.

Another strategy is to take the time to explain new studies so that community members have a high, if non-technical, level of understanding of the risks and benefits. The scientists introduce research ideas to community leaders and then present them at wider gatherings so that the public can ask questions and decide whether to accept the study. As the meetings go on, Mharakurwa says, some community members end up answering other members' questions. "Then you know your study is accepted," he says.

Timing the introduction of new studies to the community is tricky. "The earlier, the better to involve the community, including in the design stage," Mharakurwa says. However, he adds, once a programme has been introduced, it must then materialize in a reasonable time frame, or else credibility might be lost and make future approaches more difficult.

Once a study is under way, it is also important for scientists to communicate to participants the benefits of getting involved, says Blessing Ahiante, a PhD researcher in the Hypertension in Africa Research Team (HART) at North-West University in Potchefstroom, South Africa. Ahiante does fieldwork on blood pressure, diet and genetics in rural and less educated communities in South Africa. She provides simplified educational materials for study participants to explain her research, and tries to give them nutrition-related advice that they can start to use right away, before the study itself gives results. "We've got them to eat more beets," Ahiante says, as well as to reduce their salt intake, do more exercise and to drink more water, all of which will help to manage the growing burden of non-communicable diseases, such as diabetes, in poor communities (see page S65).

> ***"False rumours can ignite if communication with the community is poor."***

ORGANIZING FOR HEALTH

The Macha Research Trust also tries to provide local people with tangible benefits from its work. For example, on the basis of satellite imagery, the centre can detect probable mosquito breeding sites and warn nearby residents by text message. Such short-term benefits encourage greater cooperation in subsequent studies, which enables the centre to do better science. For instance, it has begun to use the rapid diagnostics that Mharakurwa developed at Oxford to test for individuals carrying the malaria parasite at a dormant stage. Persuading busy people to allow researchers to test them when they were asymptomatic wasn't easy, but it might have contributed to successfully lowering malaria rates in the region.

Mharakurwa is now working to apply Macha's approach to research elsewhere, including his home country, Zimbabwe. "I believe what Macha does is really the right way of building trust with the community," he says.

"I hope he succeeds, but it's not a given," Agre says. "The governance of Zimbabwe has been unstable and the economy is severely stressed," he explains — factors that can disrupt public-health efforts.

Other health-care researchers and practitioners have experienced broken community trust for reasons that range from complacency to bad luck. In what must be the most extreme case, efforts to eradicate polio were subjected to years, if not decades, of delays in Pakistan after US intelligence operatives used health workers to assist in the hunt for al-Qaeda leader Osama bin Laden. The ploy has led to mistrust of vaccination drives and even occasional acts of violence against vaccine workers.

Events such as these have contributed to a growing awareness of the risks of research or implementation programmes, Mharakurwa says. "False rumours can ignite if communication with the community is poor," he adds.

The Macha approach is as an example of the sort of continuity and investment in human relationships that is needed if science is going to succeed in the poorest parts of the world. Solutions to malaria and many other diseases exist already, says Agre. The problem is that they are not available to the people who need them the most. The onus is on political leaders, he says, to properly allocate the resources needed — or to at least get out of the way of health-care workers. And organizations must win the co-operation of community members for both research and implementing interventions. "Organization is what improves health," Agre says, "We don't need to reach for magical cures." ■

Lucas Laursen *is a freelance journalist in Madrid.*

CHRISTIAN FLEMMING/LINDAU NOBEL LAUREATE MEETINGS

Elizabeth Blackburn gave a keynote lecture at the Lindau meeting earlier this year.

POLICY

Science as a global public good

Is it time to support and manage science as a common resource?

BY LUCAS LAURSEN

Elizabeth Blackburn's work on telomeres, for which she was jointly awarded the 2009 Nobel Prize in Physiology or Medicine, has turned her into a socially minded scientist. In a keynote lecture at the 68th Lindau Nobel Laureate Meeting in June, Blackburn — a biologist at the University of California, San Francisco — called on scientists young and old to follow the same path: "Let's use our scientific prowess to be more active, politically."

Blackburn proposed that researchers should adopt the 2015 Paris climate agreement as a model for garnering long-term, international support for science. Her vision goes beyond obtaining funding and includes goals such as increasing the geographical and ethnic diversity of the scientific community, as well as making the publishing process faster and more open. "This proposal is rather like the early stages of climate change, when one saw old ways of doing things that inadvertently led to disadvantageous and unanticipated consequence," she says.

Treating science as a local issue, managed mainly by centralized research agencies or foundations, for example, is inadequate. In a panel discussion alongside Blackburn at Lindau, Nobel laureate in chemistry Martin Chalfie bemoaned the way in which decision-making in science continues to differ from one country to the next. "There really needs to be a real thinking about global cooperation in terms of the sciences," he said.

That's because the problems that science aims to address span borders.

Blackburn's research on telomeres — DNA caps at the ends of chromosomes that prevent the loss of genetic information during cell division — had implications that made her conscious of the fact that many health problems can be solved through social change. For example, air pollution has been linked to telomere shortening, which can lead to premature cell ageing and an increased risk of disease, including certain cancers. The most straightforward fix for this issue, she suggests, would be to devise and enforce policies that aim to keep the air clean. And because particulate matter is not restricted by national boundaries, such a solution must be applied worldwide.

To craft science-based policies and ensure that they reach everyone, Blackburn says, some of the old approaches — pursuing short-term national interests and budgets, for example — are no longer good enough. During the Lindau keynote, she suggested that sharing knowledge, through open-access publishing or public research initiatives, is better than competing in walled gardens. She also said that meeting the challenges of today will require the expertise of a diverse array of researchers to generate stronger ideas.

A way to achieve this is to better use the skills of those in low- or middle-income nations. "Science misses out on developing-country talent if there's not the education to support it," she says.

Some high-income countries such as Germany have programmes that are designed to recruit researchers from poorer countries into university-level training schemes or postdoctoral positions, with the aim of supporting the next generation of scientists in those regions.

Institutions are learning to overcome other biases that have limited the diversity of people in science, says Blackburn. She points to progress towards tackling age discrimination, in particular. The US National Institutes of Health and the European Research Council, for instance, have both set aside funds for early-career researchers to ensure that they can access certain grants without having to compete with more-senior scientists, who can benefit from entrenched advantages. "It's a culture change," she says, that will only spread.

Open-access journals will ensure that once knowledge exists, more researchers will be able to use it. But such efforts will require funding that crosses borders. So it's time, Blackburn says, "to make sure that national governments show a real, serious, long-term commitment to science research". Blackburn suggests the 70th Lindau Nobel Laureate Meeting, planned for 2020, as an opportunity for signing such an accord. "I very deliberately avoided specifics, because once one opens that series of discussions — the 'how' — then the discussion can quickly get lost in the vortex of objections," Blackburn says. "It is at the early aspirational and — I hope — inspirational stage, now."

Science-based platforms for change have coalesced before. Three years ago, at the 65th Lindau Nobel Laureate Meeting, Blackburn joined 35 other Nobel laureates in signing the Mainau Declaration 2015 on Climate Change, which echoed the sentiment of its precursor on the dangers of nuclear weapons. Later that year, a delegation of signatories delivered the declaration to then-president of France François Hollande during the 2015 United Nations Climate Change Conference in Paris.

Reflecting on the idea that science can, and should, be less centralized, Blackburn says that she is looking to the more than 500 early-career researchers from around the world who attended this year's Lindau meeting, as well as their peers, to drive forward her proposed agreement on the globalization of science. She asks, "Is this an idea that can be embraced by more than just the young scientists at Lindau?" ■

Lucas Laursen *is a freelance journalist in Madrid.*

JACOB SILBERBERG/ PANOS

Peru has struggled to get a handle on the geographical distribution of non-communicable diseases throughout the nation.

HEALTH CARE

A shifting burden

Health-care providers in low- and middle-income countries are having to adapt to deal with the growing problem of non-communicable disease.

BY CHARLES SCHMIDT

In the mid-2000s, a physician named Jaime Miranda began looking into how urban migration was affecting the health of people in his native Peru. Rural villagers fleeing political violence were descending on the capital, Lima, and Miranda noticed that these urban transplants seemed prone to obesity, type 2 diabetes and other non-communicable diseases (NCDs). Miranda wanted to compare rates of NCDs in these migrants with those in other Peruvian populations. But doing so was a challenge: NCD surveys were almost non-existent in the country. "We had no good way to estimate the magnitude of the problem," Miranda says.

Inadequate capacity for NCD research — and for preventing and treating chronic diseases — is not limited to Peru. Fuelled by mass urbanization, sedentary lifestyles and the growing availability of nutrient-poor processed foods, NCDs have emerged as leading causes of death in low- and middle-income countries (see 'A weighty threat'). Cardiovascular disease, cancer, respiratory diseases and complications from diabetes now kill an estimated 15 million people a year; 85% of these premature deaths occur in low- and middle-income countries[1]. And death rates from NCDs are rising steadily in poor countries. These diseases strike people younger in poor countries than in high-income ones, and they rob fragile economies of crucial human capital.

Health scientists such as Miranda want to identify local NCD risk factors, so that they can deliver more-targeted responses. But shortfalls in capacity hobble their efforts. Health systems in poorer countries have evolved to cope with what were until recently much greater dangers to health, such as infectious diseases and acute threats to child and maternal health — not chronic conditions that can require lifelong management.

Alarm over a rising tide of NCDs is now pushing the global health community to act. The United Nations and the World Health Organization have set ambitious targets for the prevention and control of NCDs, and emerging partnerships between scientists from low- and middle-income countries are working to build the needed capacity. This is a break from an older model that had experts from high-income countries parachute in and dominate projects on the ground. Collaborators are trying to shore up existing primary care and build the epidemiological capacity to monitor and respond to chronic diseases, while providing senior mentorship that helps young scientists to identify relevant research

BRYAN MELTZ/EMORY UNIV.

Venkat Narayan (standing) specializes in studying chronic diseases in southern Asia.

questions, conduct appropriate investigations and publish results in respected journals. "Many countries still lack a culture of NCD research," says Venkat Narayan, a physician and epidemiologist at Emory University in Atlanta, Georgia. "They don't have the right skill sets and haven't had the time to get to the right questions, and to come up with appropriate study designs," says Narayan, who works on numerous chronic-disease projects in South Asia. "This is where global collaborations can really help."

THE EXPERIENCE IN PERU

Peru's poor ability to monitor chronic disease was evident to Miranda. Incidence of hypertension and diabetes was higher than ever, but prevalence rates weren't being broken down into subgroups — a major shortcoming because NCD risk factors vary between groups living in cities, the countryside and the mountains.

But then in 2008, after he had finished a PhD in epidemiology at the London School of Hygiene and Tropical Medicine, Miranda learnt of a promising opportunity: the US National Heart, Lung, and Blood Institute (NHLBI) had teamed up with insurance company UnitedHealth Group in Minnetonka, Minnesota, to establish a network of 11 NCD centres of excellence in 10 low- and middle-income countries. The centres were each tasked with building local capacity to prevent and control chronic diseases, and had to partner with an academic institution in the United States, Canada, Europe or Australia. Enticed by the chance to compete for funding that would support an entire platform for NCD research, rather than just a single project, Miranda approached his friend and former PhD-thesis adviser Robert Gilman to see if he'd join the grant application. An epidemiologist at Johns Hopkins University in Baltimore, Maryland, Gilman already had extensive experience in Peru. He'd spent nearly two decades there building up three sites dedicated to monitoring infectious disease. Located in mountainous, coastal and urban settings, each site had been outfitted with computer systems, diagnostic-testing laboratories and other resources needed to characterize disease rates in the population. Now that the health burden in Peru was shifting towards NCDs, Gilman welcomed the chance to refocus the work at the sites.

After the grant application was accepted in 2008, to the tune of US$5 million over 5 years, Miranda took the lead in establishing the CRONICAS Centre of Excellence in Chronic Diseases, based at Cayetano Heredia University in Lima. Fieldworkers started gathering health information from locals, using questionnaires and performing diagnostic examinations to confirm the presence or absence of chronic diseases and underlying risk factors. With the arrival of William Checkley, a Johns Hopkins physician and pulmonary-disease specialist, in 2009, CRONICAS added lung disease to its existing strengths in heart disease and diabetes. And from these nascent efforts, a database of NCD prevalence rates began to grow.

Now nine years old, CRONICAS has evolved into one of Latin America's leading NCD research centres — Narayan describes it as a "model of interdisciplinary research that is scarce in any part of the world" — and it continues to thrive on grants and donations. Scientists at CRONICAS have published widely in major journals, and the centre's research is now expanding into areas such as mental health and the health effects of environmental pollution. Current projects include an investigation of consumer responses to nutritional warnings on food labels and a clinical trial assessing the effectiveness of mobile-phone technology for monitoring depressive symptoms in people with hypertension and diabetes. CRONICAS builds research capacity by providing postgraduate fellowships to investigators with varied backgrounds, and inviting junior investigators to propose research projects and take the lead in grant applications. "We want to give opportunities to students so they can get a feel for what it's like to do research that's consistent with international standards," Miranda says.

CAPACITY BUILDING IN SOUTH ASIA

Thousands of kilometres away in New Delhi, the Public Health Foundation of India (PHFI) is a fast-growing powerhouse for research and teaching on NCDs. Dorairaj Prabhakaran, a cardiologist and vice-president of the PHFI, says that South Asian countries still have a long way to go in terms of building not just the research capacity but also the clinical capacity needed to prevent and treat chronic disease. For instance, like many other low- and middle-income countries, India has an acute shortage of primary-care doctors with adequate experience in NCDs. Furthermore, chronic diseases "aren't given their due" in South Asian medical schools, Prabhakaran says.

To augment clinical capacity, the PHFI has launched a training programme on NCDs for practising physicians. More than 21,000 physicians have so far taken its year-long courses, which have now been expanded into Nepal, Myanmar, Bangladesh, Afghanistan and several dozen countries in Africa. Classes convene once a month, and cover wide-ranging topics, such as cardiovascular disease, stroke, women's health and diabetes. The course augments teaching in medical school by providing "updated, focused and in-depth training in specific chronic conditions", says Arun Jose, a programme manager at the PHFI.

But providing additional training for doctors is just part of the PHFI's broader strategy. The foundation also conducts training programmes for community health workers, who increasingly are helping to ease clinical burdens by acting as first points of contact between individuals and their local health-care systems. According to Jose, community health workers dispense lifestyle advice, assist with patient screening and physician referrals and assume responsibilities for follow-up and rehabilitative services. And their efforts are increasingly technology-enabled. For instance, the PHFI and its collaborators have created low-cost technology for monitoring people with poorly controlled hypertension and type 2 diabetes. To use it, nurses and community health workers enter patient information into a mobile-phone-based

clinical-support system. The software collects and stores patient data, and generates doctor-vetted treatment recommendations that are based on each person's health status and past medical history[2]. The system also sends text messages to remind patients to keep up with their treatments. A study of the technology showed improvements in blood pressure, glucose control and other clinical outcomes in patients in resource-poor settings.

One of the PHFI's major contributions has been to establish the ongoing study of 28,000 people from Delhi and Chennai in India, and Karachi in Pakistan, who undergo annual screening for heart disease and type 2 diabetes. Data generated by this work, collated by the foundation's Center for Cardiometabolic Risk Reduction in South Asia (CARRS), which was spawned by the same grant programme as CRONICAS, have become a valuable resource for researchers investigating how lifestyle and chronic-disease risks vary at the individual and community level. On scouring CARRS data, investigators have found that chronic diseases in South Asia don't always present as they do in more-affluent countries. Heart disease, for instance, often presents at a younger age in CARRS participants and in the absence of typical risk factors such as high blood pressure. Prabhakaran points out that participants who were malnourished early in life are prone to developing metabolic disease — a finding that's consistent with results from Miranda's research on urban migrants in Peru.

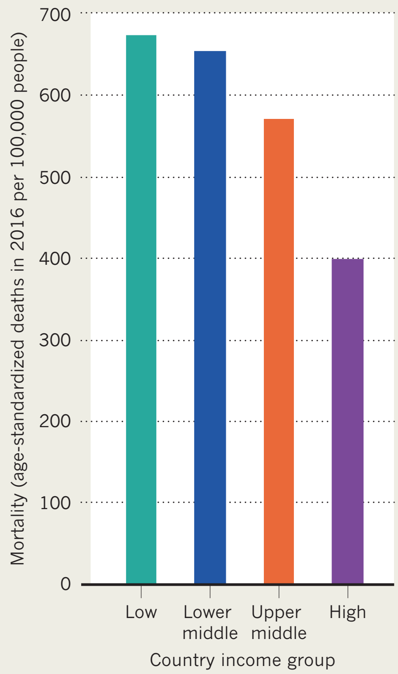
While analysing CARRS data, Lisa Staimez, a public-health researcher at Emory, also discovered[3] that participants from the Chennai study site were prone to type 2 diabetes, despite being very thin. This runs counter to the pattern in the West, where the disease is more often associated with obesity. This 'lean' diabetes arises from poorly functioning β-cells that don't produce enough insulin to meet physical needs, Staimez says, and it has since been documented throughout South Asia and sub-Saharan Africa. Importantly, this form of type 2 diabetes can't be remedied by standard interventions for the disease, Staimez says, such as weight loss or treatment with metformin (a drug that lowers blood sugar). "We need more research capacity to figure out why we have these differential chronic-disease patterns," says Narayan, who collaborated with Staimez on the research. "Otherwise, clinicians might believe, naively, that we should treat these conditions in developing countries the same way we do here in the United States or in Europe."

"He built an itinerary that sent me to all the hotspots where the disease was occurring."

Building the capacity to intervene early in people with type 2 diabetes is important because, left untreated, the condition spirals towards kidney disease. Chronic kidney disease is one of the fastest growing causes of death worldwide — in many countries, it now ranks in the top five. In India, rates of the illness are thought to have grown by 50% since 2003. But kidney-disease surveillance is dismal in India and in other low- and middle-income countries, and there is a pressing need to find and treat people before they need life-long dialysis or kidney transplants — both exceedingly expensive. People with diabetes make obvious candidates for screening. According to Shuchi Anand, a nephrologist at Stanford University School of Medicine in California, around half the people affected by kidney disease in developing countries have diabetes. The rest have kidney disease resulting from non-conventional risk factors that remain poorly understood, such as low birth-weight, chronic dehydration or inherited gene defects. Anand and Prabhakaran are now analysing blood and urine samples obtained from the CARRS cohort to look for markers that might predict kidney disease in people without conventional risk factors. "Ideally, they'll point to the high-risk people we should be screening routinely," Anand says.

A WEIGHTY THREAT

Rates of death due to non-communicable diseases are greater in low- and middle-income countries than in high-income nations when differences in the age structures of their respective populations are taken into account.



SOURCE: WHO GLOBAL HEALTH ESTIMATES

FRIENDLY COLLABORATIONS

Anand says that it's crucial to network with in-country partners who understand local constraints on NCD research and how to overcome them. Much of her fieldwork on chronic kidney disease of unknown origin has taken place in Sri Lanka, where she says local nephrologist Nishantha Nanayakkara has provided indispensable assistance. Nanayakkara works at Kandy Teaching Hospital in Kandy, as well as in a bare-bones clinical facility located in the remote city of Girandurukotte. "He built an itinerary that sent me to all the hotspots where the disease was occurring," Anand says. "I could never have done without his help." While Nanayakkara continues to help Anand survey chronic kidney disease in the country, Anand is returning the favour by providing technical assistance to medical officers at Kandy on the use of peritoneal dialysis — a home-based treatment that filters and cleans blood inside the body, rather than through an external machine. Many Sri Lankans with chronic kidney disease of unknown origin live in rural areas, and travelling two or three times a week for dialysis is difficult because it limits the time they can spend working. Peritoneal dialysis could increase the number of patients under a doctor's care, Anand says, but both patients and carers need significant training on how to use the equipment.

Salim Yusuf, a cardiologist at McMaster University in Hamilton, Canada, has worked on NCD projects in more than 100 countries, and says that the most productive collaborations happen when partners generate research questions together. "What's most important is having a committed investigator on the ground," he says. "It's the personal touch that makes the difference." Miranda agrees. A crucial factor behind the success of CRONICAS, he says, is that the way it was funded required the investigators in Peru to manage the grant. "We had to mature scientifically and bring our administrative and grant management procedures up to international standards quickly," he says.

With threats from NCDs rising steadily, the need to address them is becoming more urgent. In 2015, the UN called for action to reduce premature mortality from NCDs by 30% by 2030. But last year, the UN secretary-general concluded that developing countries still lack the capacity needed to develop and implement appropriate response strategies. The burden of NCDs falls disproportionately on the poorest nations, but most of the research devoted to these conditions takes place in richer countries. Narayan says that points to a lost opportunity. "We live in an interconnected world," he says. "And there are huge discoveries waiting to be made by expanding collaborative research. It's the right thing to do." ■

Charles Schmidt *is a science writer based in Portland, Maine.*

1. World Health Organization. *Noncommunicable diseases* http://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (2018).
2. Ajay, V. S. *et al. J. Am. Heart Assoc.* **5**, e004343 (2016).
3. Staimez, L. R. *et al. Diabetes Care* **36,** 2772–2778 (2013).

TAJ FRANCIS

SCIENCE WITHOUT BORDERS

4 BIG QUESTIONS

Working as a scientist in low- and middle-income countries can be challenging, but it also provides the opportunity to make a difference to people's lives.

BY RICHARD HODSON

	QUESTION	WHAT WE KNOW	WHAT WE CAN DO	THE RESEARCHER'S VIEW
1	***How can scientists in training best prepare for a research career in Africa?***	Budding researchers often study in other countries. When they return to Africa, many find themselves facing unexpected challenges — from coping with an overwhelming amount of teaching to having to find their own funding.	Programmes are emerging that provide scientists with financial and technical support to establish a career in Africa. With the help of local mentors, students can plan for the challenges ahead, for example, by choosing a specialism that is likely to attract funding.	"You can get all the training you want overseas, but it doesn't mean that when you come home the environment is conducive, or that you are actually prepared." **Mashiko Setshedi**, *Groote Schuur Hospital, South Africa*.
2	***How can health-care systems in low- and middle-income nations adapt to changing needs?***	Non-communicable diseases such as diabetes and cardiovascular disease are on the rise in low- and middle-income countries, claiming the lives of millions each year. Health systems in many of these nations lack the capacity to monitor, prevent or treat such conditions.	Infrastructure can be repurposed for non-communicable disease research and health workers can be trained to better understand chronic diseases. International researchers need to collaborate with in-country scientists to take advantage of their local knowledge.	"Many countries still lack a culture of non-communicable-disease research. They don't have the right skill sets and haven't had the time to get to the right questions". **Venkat Narayan**, *Emory University, Atlanta, Georgia*.
3	***What can scientists do to gain and maintain the trust of local communities?***	Distrust of researchers and health workers in local communities can limit scientist's ability to do their work. A lack of trust and understanding, particularly during disease outbreaks, has hampered efforts to contain disease, and even led to violence.	Reducing the turnover of staff at research centres can engender trust through familiarity. Some centres, such as the Macha Research Trust in Zambia, also make an effort to involve members of the local community in the planning and execution of research projects.	"False rumours can ignite if communication with the community is poor." **Sungano Mharakurwa**, *Africa University, Mutare, Zimbabwe*.
4	***What can the research establishment do to pave the way for young scientists to succeed?***	For scientists who are just starting out, funding can be difficult to come by — more-established researchers are often favoured. Some scientists can feel as though they are just an extra pair of hands for their supervisor.	Several funding agencies are setting aside portions of money for early-career scientists. Governments of some high-income countries, including Germany, offer training specifically for scientists from low- and middle-income countries.	"We are doing a lot of damage to our science by not supporting young people." **Michael Levitt**, *Stanford University, California*.

Richard Hodson *is supplements editor at* Nature.

ILLUSTRATION BY MARCO MELGRATI

Strength in isolation

Deep local collaborations and research with regional character are powerful drivers in Western Australia, one of the world's most remote scientific communities.

BY JACK LEEMING

The largest of Australia's six states, Western Australia covers about one-third of the country's landmass, and is home to 11% of its population. Almost three-quarters of residents live in the coastal city of Perth, the state capital. The next-nearest large city is Adelaide, a three-hour flight away in the neighbouring state of South Australia.

Sometimes, says George Yeoh, this isolation makes his life difficult. "If we go over east, it's probably a couple of days," says Yeoh, a liver researcher who runs the Centre for Cell Therapy and Regenerative Medicine at Perth's Harry Perkins Institute of Medical Research. "When I invite people from Melbourne or Sydney to give a talk, they say, 'Sorry, George, I made other commitments; I can't fly in in the morning and fly back the same day.'"

All of this gives Western Australia in general, and its research scene especially, a sense of independence. "We tend to think a little different from the folks on the east seaboard," says Yeoh.

Perth sprawls inland from the coast for approximately 50 kilometres, from a steel-and-glass city centre around the Swan River to well-spaced rows of garden-flanked dwellings, most of them bungalows, which give way first to crop fields and then to a red–orange rusty brush scratched with dirt tracks. Much of the state's 2.6 million square kilometres is sparsely populated and has rarely been visited — except by Indigenous Australians who arrived on the continent at least 65,000 years ago.

Across this vastness, Western Australia's outback is studded with plentiful reserves of precious metals and minerals. When gold was discovered in the 1890s in Kalgoorlie, now a seven-hour drive from Perth, the resulting economic boom transformed Perth from a small provincial town into Australia's fourth most-populous city.

Much of the state's economy — and science — is dominated by the resources sector, which covers the exploration, extraction and processing of minerals and hydrocarbons. Curtin University is one of five universities in Western Australia, four of which are in Perth. It is a world leader in mining technology and research.

ISOLATION

Many west-coast scientists feel that their work struggles to hold the attention of funding agencies and the federal government in Canberra. Researchers suspect that the government prefers the nearer and larger cities and the biotech hubs of the east coast.

"I tell my students that is something we need to overcome," Yeoh says.

"We have to be very collaborative here,"

agrees Sue Fletcher, a biomedical researcher at Murdoch University, also in Perth. "We're a smaller city. We don't have the same number of institutions with big expensive equipment, so the universities work together really well to access each others' facilities."

Data support this perspective. A 2016 report by Nature Index, which tracks scientific-article author affiliations, found that Curtin University and the University of Western Australia (UWA) formed Australia's strongest collaboration between two research institutes, coming 73rd globally. No other Australian university made it into the list of the top 100 partnerships. The data also showed that Curtin was the most collaborative Australian university overall.

"Collaboration comes naturally here," says Robyn Owens, deputy vice-chancellor for research at UWA. "It's only by doing that that we can ensure we have the appropriate research infrastructure in Western Australia."

Peter Klinken, the state's chief scientist, says scientists in Western Australia must be proactive if they are to collaborate farther afield. "From my perspective, don't whinge about it. Get out and do something about it. You have to be prepared to do a two-day trip. A lot of decisions are made in meeting rooms or in corridors or over a cup of coffee. We don't have that luxury here."

Klinken is an energetic man, with radical plans for how science could shape Western Australia. Twenty minutes into an interview with *Nature* at his office in Perth, he spins out of his chair to gesticulate at a whiteboard bearing multicoloured scrawls indicating specific areas of interest: "Li", "Automation", "Biodiversity".

Klinken has a vision for encouraging even deeper collaborations between Western Australia's universities, by merging them into a single institute, in the style of the University of California system or the federated universities of London. "I have one suggestion: that all of the universities are called the University of Western Australia and there are four campuses," he says. "It's a way of getting greater coordination." Klinken acknowledges that he "was met with stony silence" when he first suggested this to university executives. But, he says, "I have raised it seriously at the highest levels of the state. And the universities are established under state acts of parliament."

He argues that a single, larger, more prestigious university would also help to draw more international students to the state — international education is Australia's third-largest export, in terms of dollars, and Western Australia provides only 8% of the country's total. "It could be a way of addressing that," Klinken says, adding that a similar merger is being discussed between the University of Adelaide and the University of South Australia.

EXTRACTIVE ECONOMY

The two tallest office buildings in Perth stand next to each other on the city's waterfront, dominating both the skyline and the Western Australian economy. They are mainly tenanted by mining giants BHP and the Rio Tinto group. Together, the companies reported profits of US$15 billion last year.

Mining directly accounted for 29% of the state's gross product last year. Some 85% of the state's merchandise exports in 2017 were in minerals and petroleum. Around half of the state's mining exports go to China.

This reliance on the resources sector means "Perth suffers from a cyclical boom-and-bust existence", says Justin Brown, a geologist who founded and runs the mining company Element 25. "The heartbeat of Perth rises and falls on the health of the resource industry."

Perth's heartbeat is especially arrhythmic because very little is done to the ore in Western Australian itself. "We dig up big rocks, turn them into little rocks, load them on a ship and sell them," says Klinken.

> I THINK THIS IS A UNIQUE MOMENT. IF **WE GRASP IT,** THIS STATE WILL REALLY **TAKE OFF.**

Element 25 has been trying something new to break that cycle, by partnering with research organizations to process manganese from ore deposits that the company discovered in 2010, about 1,000 kilometres north of Perth. The discovery is the largest onshore manganese resource in Australia: enough, if it was all converted into metal, to meet the world's total demand for the resource for an entire year.

Under normal circumstances, Brown's manganese deposit would have been dug up and exported to China to be converted into steel for the country's construction industry, in a toxic, energy-hungry process. Ninety per cent of the manganese that passes through human hands goes into industrial steel alloys.

Just when Element 25 would normally have been arranging a sale of its ore to China, however, the effects of the global recession brought the world's construction sector to a standstill. In 2013, iron-ore sales alone brought around Aus$75 billion (US$46 billion) to Western Australia. Two years later, that figure had dropped to less than Aus$50 billion. Demand for manganese slumped.

To save costs and pursue a manganese market not involved in construction, Element 25 partnered with the Mineral Resources unit at Australia's national Commonwealth Scientific and Industrial Research Organisation to develop a process that can leach 95% of manganese from ore at room temperature and pressure within 30 minutes — without any of the dirty processes used in China's steel production.

The partnership was made possible by a grant from the federal government, designed to link industry and research, of Aus$100,000. This was matched and then added to by the company. "We have a much cleaner, cheaper leaching step," Brown says. "It's really fun science."

NATURAL RESOURCES

Australia's reliance on natural resources is a hot political issue. Last month, the country's new prime minister, Scott Morrison, abandoned Australia's emissions-reduction policy, and thus effectively its commitment to the 2015 Paris climate agreement, making Australia the second developed country to do so after the United States. The move came during a period of drought and wildfires along the eastern seaboard, and in spite of a public poll suggesting that 70% of Australians want the government to end the country's dependence on coal.

The effects of those sentiments are being felt by Curtin's Mineral and Mining Engineering programme, the second best in the world behind the Colorado School of Mines in Golden, according to the 2018 QS World University Rankings. Undergraduate student Anis McGowan says that the resources industry is facing a recruitment crisis owing to environmental concerns about mining in Australia and because of the tough working conditions in many mines. She finds this frustrating: environmental impact aside, mining is still an essential means of satisfying the world's demand for resources. "If it's not grown, it's mined," she says, shrugging.

Sam Spearing, who runs Curtin's school of mines at the Kalgoorlie campus, where McGowan is studying, says, "We're missing around two-thirds [of our usual undergraduate cohort] this year. It's bad."

ECOLOGICAL COSTS

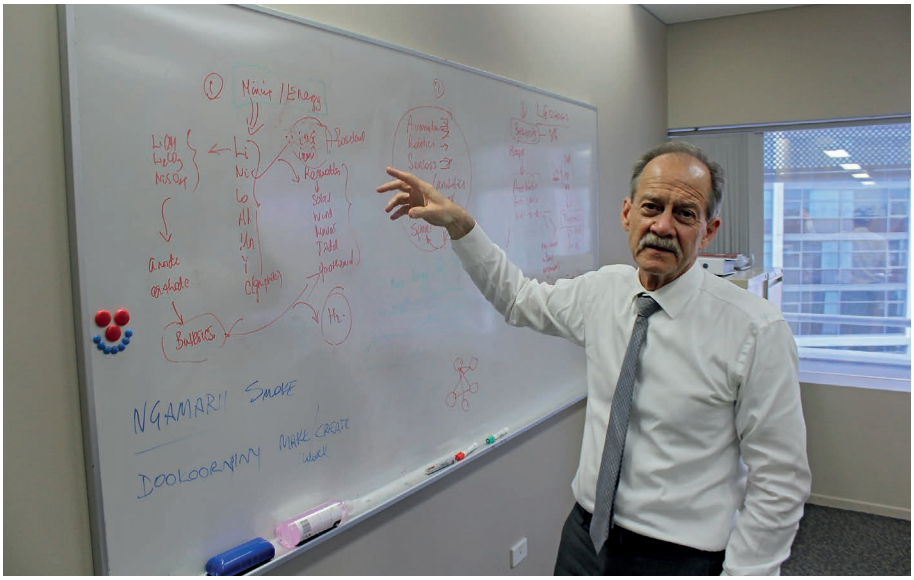
One source of opposition to Western Australia's mining sector comes from another unique element of the state's geography: its flora and fauna.

Roberta Bencini, a fauna ecologist at UWA, says that the area's isolation by vast deserts, and the even vaster ocean surrounding Australia, has given rise to a huge diversity of species found almost exclusively in Western Australia. These include exotically named and severely threatened creatures, such as the dibbler (*Parantechinus apicalis*), quokka (*Setonix brachyurus*), chuditch (*Dasyurus geoffroii*) and noisy scrub-bird (*Atrichornis clamosus*) (respectively, a small nocturnal carnivore, a marsupial with a seemingly permanent smile, a cat-sized carnivore and an aptly named bird).

"We have unique plants, unique animals," says Bencini, "but we have an array of habitat destruction, creating a big issue."

The short gestation periods of marsupials — the dominant group of mammals on the

JACK LEEMING/NATURE



Peter Klinken, Western Australia's chief scientist, has radical plans for science's role in shaping the state.

continent — make them especially vulnerable to anthropogenic changes, because populations can take much longer to recover from harm than do mammal populations elsewhere in the world.

Take the tree-dwelling western ringtail possum (*Pseudocheirus occidentalis*), for example. "It's an arboreal mammal — it's restricted by an area where there's a continuous canopy. They don't like coming to the ground," Bencini explains.

Because of this, a badly routed highway can cut local habitats in half. "There's been a 99% reduction in their inland population," says Bencini, who five years ago oversaw a successful project to build a rope bridge over one road, melding a possum habitat back together.

Another thing that's been cut too aggressively is funding for ecology research, adds Bencini. "We've been working on a shoestring budget," she says. "Instead of collaborating, we're competing for grants and funding; it's becoming cut-throat. It's very dire at the moment."

STARGAZING

Western Australia's geography is essential to another scientific field: astronomy. The central area of the vast state is largely uninhabited, with little radio coverage — making it perfect for detecting faint radio signals from elsewhere in the Universe.

"It's radio quiet here," says Melanie Johnston-Hollitt, an astronomer who runs the Murchison Widefield Array at Curtin. "It's flat, it's dry most of the time and it's got this fantastic view of the sky — there's nothing there." However, the Western Australian deserts' frequent lightning storms can occasionally affect individual units in the array.

The Murchison project is an international collaboration to study and map the wider Universe in low-frequency radio waves. It is also being used as a test bed for various technologies that might be adopted by the Square Kilometre Array — a much larger proposed project that, if funded, will be built in South Africa and Western Australia.

MOVE AWAY FROM COAL

But there are signs that Western Australia's economy is diversifying. Lithium — a resource of which the state has deep reserves — is in growing demand as the world hopes to manufacture more lithium-ion batteries. Instead of selling the ore, Western Australian policymakers are hoping to position the state as a battery manufacturer.

To back up this change of heart, Klinken is fast to point out that the state's Labor government has made science and technology a major part of its latest policy document.

For Klinken, renewable energy offers the state an opportunity. "We've got great sunlight here — more than anywhere else in the world. It's [got] the second windiest capital in the world, after Wellington," he says. "We have an opportunity to make a transition away from coal.

"All of the energy companies are saying, 'What's next?' I haven't seen a change in mindset like I've seen in the last couple of years from these big energy players. It's moving at a pace I absolutely did not expect," Klinken says. "The world is changing. You can feel it."

"If I was doing a PhD, I would be more optimistic of prospects in the future here," says Yeoh.

For Klinken, this isn't enough. "I think this is a unique moment. If we grasp it, this state will really take off. If we don't, I shake my head at what my kids and grandkids will say to me. There's something about this time here that says we really need to grasp the nettle." ■

Jack Leeming *is Spotlight and careers community editor at* Nature.